SoK: Chasing Accuracy and Privacy, and Catching Both in Differentially Private Histogram Publication

TRANSACTIONS ON DATA PRIVACY 13 (2020) 201–245

SoK: Chasing Accuracy and Privacy, and Catching Both in Differentially Private Histogram Publication

Boel Nelson[*†], **Jenni Reuben**[*‡]

[†]Department of Computer Science and Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden.

[‡]Department of Mathematics and Computer Science, Karlstad University, SE-651 88 Karlstad, Sweden.

E-mail: boeln@chalmers.se, jenni.reuben@kau.se

Received 7 March 2020; received in revised form 12 October 2020; accepted 15 December 2020

Abstract. Histograms and synthetic data are of key importance in data analysis. However, researchers have shown that even aggregated data such as histograms, containing no obvious sensitive attributes, can result in privacy leakage. To enable data analysis, a strong notion of privacy is required to avoid risking unintended privacy violations.

Such a strong notion of privacy is *differential privacy*, a statistical notion of privacy that makes privacy leakage quantifiable. The caveat regarding differential privacy is that while it has strong guarantees for privacy, privacy comes at a cost of accuracy. Despite this trade-off being a central and important issue in the adoption of differential privacy, there exists a gap in the literature regarding providing an understanding of the trade-off and how to address it appropriately.

Through a systematic literature review (SLR), we investigate the state-of-the-art within accuracy improving differentially private algorithms for histogram and synthetic data publishing. Our contribution is two-fold: 1) we identify trends and connections in the contributions to the field of differential privacy for histograms and synthetic data and 2) we provide an understanding of the privacy/accuracy trade-off challenge by crystallizing different dimensions to accuracy improvement. Accordingly, we position and visualize the ideas in relation to each other and external work, and deconstruct each algorithm to examine the building blocks separately with the aim of pinpointing which dimension of accuracy improvement each technique/approach is targeting. Hence, this systematization of knowledge (SoK) provides an understanding of in which dimensions and how accuracy improvement can be pursued without sacrificing privacy.

Keywords. accuracy improvement, boosting accuracy, data privacy, differential privacy, dimensionality reduction, error reduction, histogram, histograms, noise reduction, sensitivity reduction, synthetic data, SLR, SoK, systematic literature review, systematization of knowledge, taxonomy, utility improvement

[*]Both the authors contributed substantially, and share first authorship. The names are ordered alphabetically.

1 Introduction

Being able to draw analytical insights from data sets about individuals is a powerful skill, both in business, and in research. However, to enable data collection, and consequently data analysis, the individuals' privacy must not be violated. Some strategies [1–3] for privacy-preserving data analysis focus on sanitizing data, but such approaches require identifying sensitive attributes and also does not consider auxiliary information. As pointed out by Narayanan and Shmatikov [4], personally identifiable information has no technical meaning, and thus cannot be removed from data sets in a safe way. In addition to the difficulty in modeling the extent of additional information that an adversary may possess from public sources in such data sanitizing approaches, the privacy notion of such approaches is defined as the property of the data set. However, it is proved in [5] that for essentially any non-trivial algorithm, there exists auxiliary information that can enable a privacy breach that would not have been possible without the knowledge learned from the data analysis. Consequently, a strong notion of privacy is needed to avoid any potential privacy violations, while still enabling data analysis.

Such a strong notion of privacy is *differential privacy* [6] (Section 2), in which the privacy guarantee is defined as the property of the computations on the data set. Differential privacy is a privacy model that provides meaningful privacy guarantees to individuals in the data sets by quantifying their privacy loss. This potential privacy loss, is guaranteed independently of the background information that an adversary may possess. The power of differential privacy lies in allowing an analyst to learn statistical correlations about a population, while not being able to infer information about any one individual. To this end, a differential private analysis may inject random noise to the results and these approximated results are then released to the analysts.

Differential privacy has spurred a flood of research in devising differentially private algorithms for various data analysis with varying utility guarantees. Given a general workflow of a differentially private analysis, which is illustrated in Figure 1, we have identified four *places* (labeled A, B, C and D) for exploring different possibilities to improve accuracy of differential private analyses.

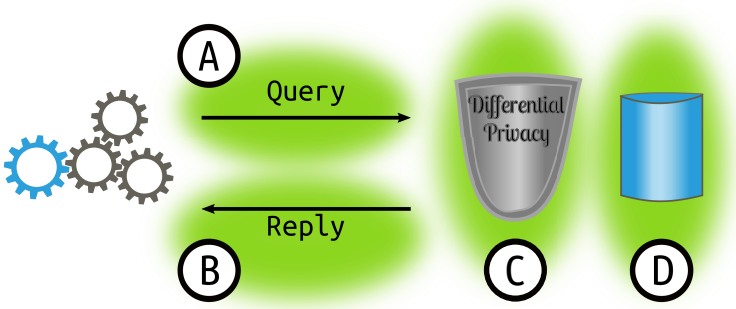

Figure 1: Places for accuracy improvement: A) Altering the query, B) Post-processing, C) Change in the release mechanism, D) Pre-processing.

In this work, we focus specifically on differentially private algorithms for histograms, and synthetic data publication. Histograms and synthetic data are particularly interesting

because they both provide a way to represent summary of an underlying data set, thus may enable further analytical tasks executed over the summary of the data set. While, histograms represent a graphical summary of frequency distribution of values of a specific domain in a data set, synthetic data is an approximate representation of data distribution of an underlying data set. Intrigued by the idea that there exists several ways to improve accuracy of privatized histograms and synthetic data without compromising privacy, we aim to systematically synthesize the state-of-the-art.

Advancement in research in differentially private histogram and synthetic data publication has received considerable interest within the computer science and statistics research communities [7–9]. However, only a few works systematically and critically assess the state-of-the-art differentially private, accuracy improving algorithms for releasing histograms or synthetic data. Li et al. [8] and Meng et al. [9] categorized different differentially private publication techniques for both histogram as well as synthetic data, and solely histograms respectively. However, their selection and categorization of the algorithms are not systematic. Further, the selected algorithms in their work are not exclusively accuracy improving techniques, but rather differentially private release mechanisms for histogram and synthetic data. That is, some of the surveyed algorithms do not boost the accuracy of an existing release mechanism by adding a modular idea, but instead invent new, monolithic algorithms. For example, some of the algorithms have discovered ways to release data that previously did not have a differentially private way of being released. Bowen and Liu [7], on the other hand, used simulation studies to evaluate several algorithms for publishing histograms and synthetic data under differential privacy. Their aim is quite different from ours, is to assess the accuracy[1] and usefulness[2] of the privatized results.

Consequently, to bridge the knowledge gap, the present paper aims to provide a systematization of knowledge concerning differentially private accuracy improving methods for histogram and synthetic data publication. To this end, we first review the main concepts related to differential privacy (Section 2), which are relevant to the qualitative analysis of state-of-the-art accuracy improving techniques for differentially private histogram and synthetic data publication (Section 5). However, before focusing on the qualitative analysis, we present our method to conduct a systematic review of literature that enable a methodological rigor to the results of the qualitative analysis (Section 3) and a review of general characteristics of the identified accuracy improving techniques (Section 4). We further study the composability of accuracy improvement techniques within the constraints of differential privacy in relation to the results of our analysis in order to pave the way for future research (Section 6). Overall, this systematization of knowledge provides a conceptual understanding of enhancing accuracy in the light of privacy constraints (Section 7).

Our main contributions are:

1. A technical summary of each algorithms in order to provide a consolidate view of the state-of-the-art (Section 4).

2. Categorization that synthesize the evolutionary relationships of the research domain in differential privacy for histogram and synthetic data publication (Section 5.1).

3. Categorization of the state-of-the-art, which is based on the conceptual relationships of the identified algorithms (Section 5.2).

[1] We will use the terms accuracy and utility interchangeably when we refer to decreasing the error, i.e the distance between the privatized result and the true results.

[2] We use the term usefulness to refer to the impact of the privatized results to conduct statistical inferences.

2 Differential Privacy

Differential privacy [6] is a statistical definition that enables privacy loss to be quantified and bounded. In differential privacy, privacy loss is bounded by the parameter ε. To achieve trivial accuracy improvement, ε can be tweaked to a higher value, as this gives less privacy (greater privacy loss) which means more accuracy. In this paper we only consider accuracy improvements in settings where ε is fixed.

We formally define ε-differential privacy in Definition 1, based on Dwork [5]. The parameter ε is usually referred to as the *privacy budget*. Essentially, ε is the cost in terms of privacy loss for an individual participating in an analysis.

Definition 1 (ε-Differential Privacy)**.** A randomized algorithm f' gives ε-differential privacy if for all data sets D_1 and D_2, where D_1 and D_2 are neighboring, and all $\mathcal{S} \subseteq \text{Range}(f')$,

$$\Pr[f'(D_1) \in \mathcal{S}] \leq e^{\varepsilon} \times \Pr[f'(D_2) \in \mathcal{S}]$$

A relaxed version of differential privacy is (ε, δ)-differential privacy Dwork et al. [10], which we define in Definition 2. (ε, δ)-differential privacy is primarily used to achieve better accuracy, but adds a subtle, probabilistic dimension of privacy loss. (ε, δ)-differential privacy is sometimes also called *approximate differential privacy* [11].

Definition 2 ((ε, δ)-Differential Privacy)**.** A randomized algorithm f' is (ε, δ)-differentially private if for all data sets D_1 and D_2 differing on at most one element, and all $\mathcal{S} \subseteq \text{Range}(f')$,

$$\Pr[f'(D_1) \in \mathcal{S}] \leq e^{\varepsilon} \times \Pr[f'(D_2) \in \mathcal{S}] + \delta$$

Theoretically, in ε-differential privacy each output is *nearly* equally likely and hold for *any* run of algorithm f', whereas (ε, δ)-differential privacy for *each pair* of data sets (D_1, D_2) in extremely unlikely cases, will make some answer much less or much more likely to be released when the algorithm is run on D_1 as opposed to D_2 [12]. Still, (ε, δ)-differential privacy ensures that the absolute value of the privacy loss is bounded by ε with probability at least 1-δ [12]. That is, the probability of gaining significant information about one individual, even when possessing all other information in the data set, is at most δ.

To satisfy differential privacy, a randomized algorithm perturbs the query answers to obfuscate the impact caused by differing one element in the data set. Such perturbation can for example be introduced by adding a randomly chosen number to a numerical answer. Essentially, the maximum difference *any* possible record in the data set can cause dictates the magnitude of noise needed to satisfy differential privacy. This difference is referred to as the algorithm's L_1 sensitivity, which we define in Definition 3, based on Dwork et al. [6].

Definition 3 (L_1 Sensitivity)**.** The L_1 sensitivity of a function $f : D^n \to \mathbb{R}^d$ is the smallest number Δf such that for all D_1, $D_2 \in D^n$ which differ in a single entry,

$$\|f(D_1) - f(D_2)\|_1 \leq \Delta f$$

Since differential privacy is a property of the algorithm, as opposed to data, there exists many implementations of differentially private algorithms. Thus, we will not summarize all algorithms, but instead introduce two early algorithms that are common building blocks, namely: the Laplace mechanism [6] and the Exponential mechanism [13].

We define the Laplace mechanism in Definition 4, based on the definition given by Dwork [14]. The Laplace mechanism adds numerical noise, and the probability density function is centered around zero, meaning that noise with higher probability (than any other specific value) will be zero.

Definition 4 (Laplace mechanism)**.** For a query f on data set D, the differentially private version, f', adds Laplace noise to f proportional to the sensitivity of f:

$$f'(D) = f(D) + Lap(\Delta f/\varepsilon)$$

Furthermore, we define the Exponential mechanism (EM) in Definition 5 based on the definition given by McSherry and Talwar [13]. The intuition behind EM is that the probability of not perturbing the answer is slightly higher than perturbing the answer. EM is particularly useful when Laplace does not make sense, for example when queries return categorical answers such as strings, but can also be used for numerical answers. The reason EM is so flexible is that the utility function can be replaced to score closeness to suit the given domain.

Definition 5 (Exponential mechanism (EM))**.** Given a utility function $u : (D \times R) \to R$, and a data set D, we define the differentially private version, u':

$$u'(D,u) = \left\{ \text{return } r, \text{where } r \text{ ranges over } R, \text{ with probability } \propto exp\frac{\varepsilon u(D,r)}{2\Delta u} \right\}$$

The semantic interpretation of the privacy guarantee of differential privacy rests on the definition of what it means for a pair of data sets to be neighbors. In the literature, the following two variations of neighbors are considered when defining differential privacy: unbounded and bounded.

Definition 6. Let D_1 and D_2 be two data sets where D_1 can be attained by adding or removing a single record in D_2. With this notion of neighbors, we say that we have *unbounded* differential privacy.

Definition 7. Let D_1 and D_2 be two data sets where D_1 can be attained by changing a single record in D_2. With this notion of neighbors, we say that we have *bounded* differential privacy.

Distinguishing between the definition of neighboring data sets is important, because it affects the global sensitivity of a function. The sizes of the neighboring data sets are fixed in the bounded differential privacy definition whereas, there is no size restriction in the unbounded case.

In the case of graph data sets, a pair of graphs differ by their number of edges, or number of nodes. Therefore, there exists two variant definitions in literature [15] that formalize what it means for a pair of graphs to be neighbors. Nevertheless, these graph neighborhood definitions are defined only in the context of unbounded differential privacy.

Definition 8 (Node differential privacy [15])**.** Graphs $G = (V,E)$ and $G^{'} = (V^{'},E^{'})$ are *node-neighbors* if:

$$V^{'} = V - v,$$
$$E^{'} = E - \{(v_1,v_2) \mid v_1 = v \vee v_2 = v\},$$

for some node $v \in V$.

Definition 9 (Edge differential privacy [15])**.** Graphs $G = (V,E)$ and $G^{'} = (V^{'},E^{'})$ are *edge-neighbors* if:

$$V = V^{'},$$
$$E^{'} = E - \{e\},$$

for some edge $e \in E$.

In certain settings, ε grows too fast to guarantee a meaningful privacy protection. To cater for different applications, in particular in settings where data is gathered dynamically, different *privacy levels* have been introduced that essentially further changes the notion of neighboring data sets by defining neighbors for data streams. These privacy levels are, user level privacy [16], event level privacy [17], and w-event level privacy [18].

Definition 10. We say that a differentially private query gives *user level privacy* (pure differential privacy), when all occurrences of records produced by one user is either present or absent.

Essentially, for user level privacy, all records connected to one individual user shares a joint privacy budget.

Definition 11. We say that a differentially private query gives *event level privacy*, when all occurrences of records produced by one group of events, where the group size is one or larger, is either present or absent.

With event level privacy, each data point used in the query can be considered independent and thus have their own budget.

Definition 12. We say that a differentially private query gives *w-event level privacy*, when a set of w occurrences of records produced by some group of events, where the group size is one or larger, is either present or absent. When $w = 1$, w-event level privacy and event level privacy are the same.

For w-event level privacy, w events share a joint privacy budget.

3 Method

We conducted a systematic literature review (SLR) [19] to synthesize the state-of-the-art accuracy improving techniques for publishing differentially private histograms as well as synthetic data. Systematic literature review, which, hereafter we will refer to as systematic review when describing generally, is a method to objectively evaluate all available research pertaining to a specific research question or research topic or phenomena of interest [19]. Although, the method is common in social science and medical science disciplines, the Evidence-Based Software Engineering initiative [20] have been influential in the recognition of systematic review as the method to integrate evidence concerning a research area, a research question or phenomena of interest in software engineering research. Systematic review provides methodological rigor to literature selection and synthesization as well as to the conclusion drawn as a result of the synthesization. The method consists of several stages that are grouped into three phases. The phases of systematic review are; i) planning the review, ii) conducting the review and iii) reporting the review.

Planning the review phase underpins the need for a systematic review concerning a research topic, a research question or phenomena of interest. Hence, in the planning stage, a review protocol that defines the research questions, as well as strategies for conducting the literature review is developed in order to minimize the likelihood of researcher bias in the selection of literature.

Following the specification of the search, the selection and the data synthesis strategy, the review is conducted (conducting the review phase) in an orderly manner. Thus, the first stage of the execution of a systematic review is the identification of all available literature. This stage involves the construction of search queries and identification of all relevant

scholarly databases. After the identification of literature on a given topic of interest, they need to be evaluated for relevance, which usually is determined through a set of selection criteria. The selected literature for a systematic review is generally referred to as primary studies. Then, in order to synthesize the results of the primary studies, data are extracted from each primary study for the analysis that is the final stage of the conducting the review phase.

Reporting the review involves the documentation of the systematic review process and the communication of the results of the systematic review.

In the following subsections we describe in detail, the process we undertake in our SLR. Figure 2 shows the high-level view of the processes followed in our SLR.

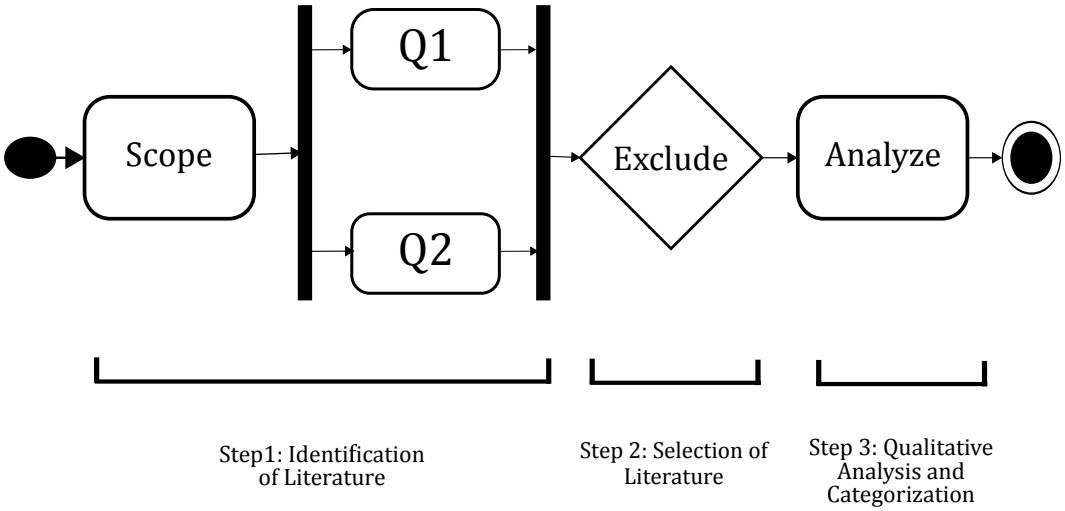

Figure 2: Workflow of processes followed in our SLR.

3.1 Identification of Literature

A thorough and unbiased search for literature is the essence of a SLR. In this SLR, we used a scholarly search engine, Microsoft Academic (MA) [21, 22], primarily for two reasons. First, for its semantic search functionality and second, for its coverage.

Semantic search leverages *entities* such as field of study, authors, journals, institutions, etc., associated with the papers. Consequently, there is no need to construct search strings with more keywords and synonyms, rather, a natural language query can be constructed with the help of search suggestions for relevant entities.

Secondly, regarding the coverage of MA. MA's predecessor, Microsoft Academic Search (MAS), suffered from poor coverage as pointed out by Harzing [23]. However, after re-launching MA its coverage has grown over the years [24, 25]. In 2017, Hug and Brändle [26] compared the coverage of MA to Scopus and Web of Science (WoS), and found that MA has higher coverage for book-related documents and conferences, and only falls behind Scopus in covering journal articles. More recently, in 2019, Harzing [27] compared the coverage of Crossref, Dimensions, Google Scholar (GS), MA, Scopus and WoS, and found that GS and MA are the most comprehensive free search engines. Accordingly, we have

chosen to use MA since it has both adequate coverage and semantic search, while for example GS lacks semantic search.

We used two queries, one focusing on histograms and the other on synthetic data. The queries are as follows, with entities recognized by MA in bold text:

Q1: **Papers** about **differential privacy** and **histograms**

Q2: **Papers** about **differential privacy** and **synthetic data**

The search was performed on June 10 2019, and yielded 159 hits in total. 78 hits for **Q1** and 81 hits for **Q2**, which are examined for relevance in the next step of the SLR process.

3.2 Selection of Literature

We constructed and followed a set of exclusion criteria Table 1 in order to select the relevant literature that provides insights to our research aim. To reflect that we specifically wanted to focus on tangible, experimentally tested algorithms, we constructed the criteria to exclude papers that contribute to pure theoretical knowledge. To select papers, we exam-

Exclude if the paper is...
1) not concerning differential privacy, not concerning accuracy improvement, and not concerning histograms or synthetic data.
2) employing workflow actions, pre-processing/post-processing/algorithmic tricks but not solely to improve accuracy of histograms or synthetic data.
3) a trivial improvement to histogram or synthetic data accuracy through relaxations of differential privacy or adversarial models.
4) concerning local sensitivity as opposed to global sensitivity.
5) not releasing histograms/synthetic data.
6) pure theory, without empirical results.
7) about a patented entity.
8) a preprint or otherwise unpublished work.
9) not peer reviewed such as PhD thesis/master thesis/demo paper/poster/extended abstract.
10) not written in English.

Table 1: List of exclusion criteria followed in our SLR.

ined the title and abstract of each paper against the exclusion criteria. When the abstract matches any one of the criteria, the paper is excluded, otherwise the paper is included. When it was unclear from the abstract that a contribution is empirical or pure theory, we looked through the body of the paper to make our decision. In the course of this stage, in order to ensure the reliability of the decision concerning the inclusion of a literature in our systematic review, both the authors have independently carried out the selection of literature stage. When comparing the decisions of the authors, if there exist a disagreement, we discussed each disagreement in detail in relation of the criteria in Table 1 and resolved it. For the full list of excluded papers along with the reason for exclusion, see Appendix A.

In the end, a total of 35 (after removing duplicates) papers were selected for the qualitative analysis.

3.3 Qualitative Analysis and Categorization

The most common framework found in the literature to analyse and understand a domain of interest, is classification schemes [28]. It concerns the grouping of objects with similar characteristics in a domain. Our aim is to synthesize; i) on the one hand, trends and relationships among each papers and ii) on the other hand, conceptual understanding of the privacy/accuracy trade-off in the differentially private histogram and synthetic data research. Therefore, from each paper we extracted distinct characteristics of the algorithms, evaluation details of the algorithms as well as design principles such as aim of the solution and motivation for using a particular technique. These characteristics are inductively analyzed for commonality, which follows, though not rigorously, the empirical-to-conceptual approach to taxonomy development defined by Nickerson et al. [28]. The categorization that resulted from the qualitative analysis are presented in Section 5.

Deviation from the systematic review guidelines in [19]: The review protocol for our SLR is not documented in the planning stage as specified by the original guidelines but rather documented in the reporting the review stage. This is largely due to the defined focus of our SLR, which is on the privacy/accuracy trade-off associated with differentially private algorithms for publishing histograms and synthetic data. Hence, the search strategy and selection criteria do not call for an iteration and an account of the changes in the process. Further, in our SLR we do not consider a separate quality assessment checklist as prescribed by the SLR guidelines. However, in our SLR the quality of the primary studies is ensured through our detailed selection criteria that involves objective quality assessment criteria for example the criterion to include only peer-reviewed scientific publications in the SLR. Furthermore, the quality of the results of our SLR is ensured through the exclusion of some of the selected primary studies because the algorithms in those studies lack comparable properties in order to perform a fair comparison with other selected algorithms. Additionally, during the analysis we surveyed additional relevant literature from the related work sections of the primary studies, which adds to the quality of the results of our SLR.

4 Overview of Papers

After analyzing the 35 included papers, 27 papers [29–55] where found to be relevant. All included papers and their corresponding algorithms are listed in the ledger in Table 2. We illustrate the chronological publishing order of the algorithms, but note that within each year, the algorithms are sorted on the first author's last name, and not necessarily order of publication.

Beware that some algorithms, for example NF, SF, have appeared in publications twice, first in a conference paper and then in an extended journal version. When a paper has two versions, we will refer to the latest version in our comparisons, but we include all references in the paper ledger for completeness. Furthermore, eight papers were excluded based on our qualitative analysis. Each decision is motivated in Section 6.2, and those eight papers hence do not appear in the paper ledger.

Furthermore, in Tables 3 and 4, we present objective parameters regarding the settings around the algorithms in each paper, for example the characteristics of the input data they operate on, and the metric used to measure errors. Our intention is that this table will allow for further understanding of which algorithms are applicable given a certain setting when one searches for an appropriate algorithm, but also to understand which algorithms are directly comparable in the scope of this SLR.

2010	Boost	Hay et al. [29]
2011	PMost, BMax	Ding et al. [30]
	Privelet, Privelet$^+$, Privelet*	Xiao et al. [31, 56]
2012	EFPA, P-HP	Ács et al. [32]
2013	NF, SF	Xu et al. [33, 57]
2014	DPCopula	Li et al. [34]
	CiTM	Lu et al. [35]
	PeGS, PeGS.rs	Park et al. [58], Park and Ghosh [36]
	DPCube	Xiao et al. [37, 59, 60]
	PrivBayes	Zhang et al. [38]
	AHP	Zhang et al. [39]
2015	RG	Chen et al. [40]
	ADMM	Lee et al. [41]
	DSAT, DSFT	Li et al. [42]
2016	(θ,Ω)-Histogram, θ-CumHisto	Day et al. [43]
	BPM	Wang et al. [44]
	PrivTree	Zhang et al. [45]
2017	DPCocGen	Benkhelif et al. [46]
	SORTaki	Doudalis and Mehrotra [47]
	Pythia, Delphi	Kotsogiannis et al. [48]
	Tru, Min, Opt	Wang et al. [49]
	DPPro	Xu et al. [50]
2018	T$^\lambda$	Ding et al. [51]
	GGA	Gao and Ma [52]
	PriSH	Ghane et al. [53]
2019	IHP, mIHP	Li et al. [54, 61]
	RCF	Nie et al. [55]

Table 2: Chronological ledger for the papers. Note that the abbreviation 'ADMM' is due to Boyd et al. [62], whereas Lee et al. [41]'s work is an extension that uses the same abbreviation.

Note that the privacy level (user, event or w-event) was not explicitly stated in most papers, in which case we have attributed the privacy level as '?'. A '?' privacy level does not imply that the algorithm does not have a particular privacy level goal, but rather, that the authors did not explicitly describe what level they are aiming for. With this notice, we want to warn the reader to be cautious when comparing the experimental accuracy of two algorithms unless they in fact assume the same privacy level. For example, comparing the same algorithm but with either user level or event level privacy would make the event level privacy version appear to be better, whereas in reality it trivially achieves better accuracy through relaxed privacy guarantees.

In general, user level privacy tends to be the base case, as this is the level assumed in *pure* differential privacy [16], but to avoid making incorrect assumptions, we chose to use the '?' label when a paper does not explicitly state their privacy level.

Given that our two queries were designed to capture algorithms that either output synthetic data or a histogram, we examine the similarity between the strategies used in each algorithm. To this end, we manually represent the similarity between the algorithms' strategies based on their output in Table 5. We distinguish between the two kinds of outputs

Ref.	Def.	Lvl.	Rel.	Dim.	In.	Mech.	Metric	Out.
[29]	ε	?	♢	1D	$\bar{x}$	LAP	MAE	HISTOGRAM
[30]	ε	?	♢	*	$\bar{x}$, $\bowtie$	LAP	MAE	CUBOIDS
[31]	ε	?	♣	1D, *	$\bar{x}$	LAP	MAE, MPE	RANGE COUNT QUERIES
[32]	ε	?	♢	1D	$\bar{x}$, $\bowtie$	LAP, EM	KL, MSE	HISTOGRAM
[33]	ε	?	♢	1D	$\bar{x}$	LAP, EM	MAE, MSE	HISTOGRAM
[34]	ε	?	♢	*, $\odot$	$\bar{x}$	LAP	MAE, MPE	SYNTHETIC DATA
[35]	(ε,δ)	ENTITY	♢	*	$\bar{x}$, $\bowtie$	MM, AGNOSTIC	MPE	MODEL
[36]	ε	?	♢	*	$\bar{x}$	DIRICHLET PRIOR	RANK CORR.	MODEL
[37]	ε	?	♢	*	$\bar{x}$	LAP	MAE	HISTOGRAM
[38]	ε	?	♣	*, $\odot$	$\bar{x}$	LAP, EM	AVD, MISS	SYNTHETIC DATA
[39]	ε	?	♢	1D	$\bar{x}$	LAP	KL, MSE	HISTOGRAM
[40]	ε	EVENT	♢	1D	$\vec{x}$, $\bowtie$	LAP	MSE	HISTOGRAM
[41]	ε	?	♢	*	$\bar{x}$	LAP, MM	MSE	CONTINGENCY TABLE, HISTOGRAM
[42]	ε	USER, *w-event*	♢	1D	$\vec{x}$, $\bowtie$	LAP	MAE, MPE	HISTOGRAM
[43]	ε	?	NODE PRIVACY	*	$\bar{x}$	EM	KS, $\ell 1$	HISTOGRAM
[44]	ε	?	♣	1D	$\bar{x}$	RR	NWSE	HISTOGRAM
[45]	ε	?	♢	*	$\bar{x}$, $\bowtie$	LAP	MPE	QUADTREE
[46]	ε	?	♢	*	$\bar{x}$, $\odot$	LAP	HELLINGER	PARTITIONING
[47]	ε	?	♢	1D	$\bar{x}$	LAP	SAQ	HISTOGRAM
[48]	ε	?	♢	1D, *	$\bar{x}$	LAP, AGNOSTIC	$\ell 2$, REGRET	*N/A*
[49]	ε	?	♢	1D	$\bar{x}$, $\bowtie$	LAP	MSE	HISTOGRAM
[50]	(ε,δ)	?	♣	*	$\bar{x}$	GAUSSIAN, MM	MISS, MSE	MATRIX
[51]	ε	?	NODE PRIVACY	*	$\bar{x}$	LAP	KS, $\ell 1$	HISTOGRAM
[52]	ε	?	♢	1D	$\vec{x}$	LAP	MAE	HISTOGRAM
[53]	ε	?	♢	*, $\odot$	$\bar{x}$, $\bowtie$	MWEM	KL, $\ell 1$	HISTOGRAM
[54]	ε	?	♢	1D,*	$\bar{x}$, $\odot$	LAP, EM	KL, MSE	HISTOGRAM
[55]	ε	?	♣	1D	$\bar{x}$	RR	MSE	HISTOGRAM

Table 3: Mapping between papers to corresponding differential privacy definition, privacy level, neighbor relationship, dimension of data, input data, use of mechanism, error metric and output data. Abbreviations and the corresponding symbols are explained in a separate table.

	Key	Meaning
Data	$\bowtie$	Correlated
	$\vec{x}$	Dynamic
	$\odot$	Sparse
	$\bar{x}$	Static
Dimension	*	Multi
	1D	Single
Mechanism	EM	Exponential mechanism
	LAP	Laplace mechanism
	MM	Matrix mechanism
	RR	Randomized response
Metric	AVD	Average Variation Distance
	KS	Kolmogorov-Smirnov distance
	KL	Kullback-Leibler divergence
	$\ell 1$	L1 distance
	$\ell 2$	L2 distance
	MAE	Mean absolute error
	MISS	Misclassification rate
	MPE	Mean percentage error
	MSE	Mean squared error
	NWSE	Normalized weighted square error
	SAQ	Scaled average per query
Relation	♣	Bounded
	♢	Unbounded

Table 4: Meaning of symbols and abbreviations.

by their different goals: for histograms, the goal is to release *one optimal histogram* for a given query, whereas for synthetic data the goal is to release a data set that is optimized for some *given set of queries*. Some algorithms use similar approaches to the algorithms from the other query; and therefore we label them as hybrid. An example of a hybrid paper is Li et al. [54], since they both deal with one-dimensional histograms (IHP), and then re-use that strategy when producing multi-dimensional histograms (mIHP) that resembles the outputs of synthetic data papers.

Histogram	Hybrid	Synthetic Data
Hay et al. [29]	Lu et al. [35]	Li et al. [34]
Xiao et al. [31]	Ding et al. [30]	Park and Ghosh [36]
Ács et al. [32]	Xiao et al. [37]	Zhang et al. [38]
Xu et al. [33]	Lee et al. [41]	Xu et al. [50]
Zhang et al. [39]	Zhang et al. [45]	
Chen et al. [40]	Benkhelif et al. [46]	
Li et al. [42]	Kotsogiannis et al. [48]	
Day et al. [43]	Wang et al. [49]	
Wang et al. [44]	Li et al. [54]	
Doudalis and Mehrotra [47]		
Ding et al. [51]		
Gao and Ma [52]		
Ghane et al. [53]		
Nie et al. [55]		

Table 5: The papers grouped by their type of output, where hybrid internally uses histogram structures where synthetic data is sampled from.

5 Analysis

We present our qualitative analysis on 27 included papers from two different perspectives in the light of research in differential privacy histogram and synthetic data. First, from a evolutionary perspective for identifying trends and to position each contribution in the history of its research (Section 5.1). Second, from a conceptual perspective for understanding the trade-off challenge in the privacy and utility relationship (Section 5.2).

5.1 Positioning

In order to provide context, we studied *where* the algorithms originated from, and how they are connected to each other. To also understand *when* to use each algorithm, and which ones are comparable in the sense that they can be used for the same kind of analysis, we also investigate which algorithms are compared experimentally in the papers.

First, we explored and mapped out the relationships between the included algorithms. To further paint the picture of the landscape of algorithms, we analyzed the related work sections to find external work connected to the papers included in our SLR. We present our findings as a family tree of algorithms in Figure 3, which addresses from where they came.

Since our exploration of each algorithms' origin discovered papers outside of the SLR's queries, we also provide a ledger for (Table 6) *external* papers. When the authors had not designated a name for their algorithms, we use the abbreviation of the first letter of all author's last name and the publication year instead. Note that we have not recursively investigated the external papers' origin, so external papers are not fully connected in the family tree.

From the family tree, we notice that there are several different lines of research present. One frequently followed line of research is that started by Xu et al. [33], NF, SF, which addresses the issue of finding an appropriate histogram structure (i.e. bin sizes) by creating a differentially private version of a v-optimal histogram. Essentially, EM is used to deter-

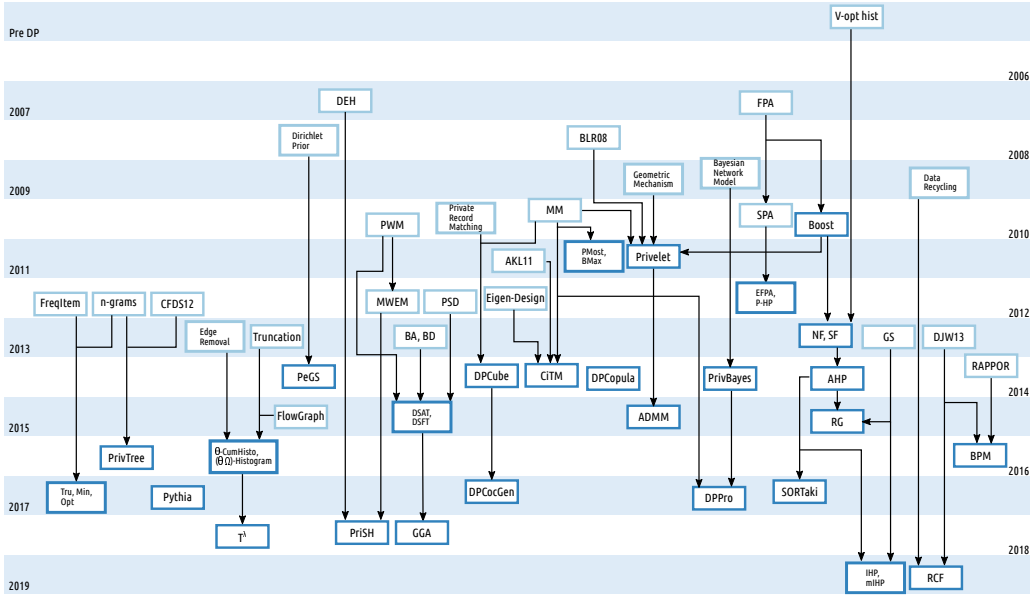

Figure 3: The family tree of algorithms. Light blue indicate papers not covered by the SLR, and the darker blue represents included papers.

mine the histogram structure, and then the Laplace mechanism is used to release the bin counts. The idea by Xu et al. [33] is followed by AHP, RG, SORTaki, IHP and mIHP.

The matrix mechanism (MM) is a building block that is used in PMost, BMax, C*i*TM and DPPro. Apart from using the same release mechanism, they do not share many similarities as also becomes apparent when comparing their experimental evaluation.

Only Pythia and DPCopula appears as orphaned nodes in the family tree. Pythia is special in the sense that it is not a standalone algorithm, but rather provides a differentially private way of choosing the 'best' algorithm for a given data set. DPCopula has a mathematical background in copula functions, which are functions that describe the dependence between multivariate variables. This approach of using copula functions is not encountered in any of the other papers.

To further put the algorithms into perspective, we explored which algorithms were used in their experimental comparisons. The comprehensive matrix of which algorithms are experimentally compared to each other in Table 7. This complements the fact table (Table 3) in addressing the question of when to use an algorithm, as algorithms that are compared experimentally can be used interchangeably for the same analysis. E.g, when NF is used, it can be swapped with for example IHP.

Label	Author
AKL11	Arasu et al. [63]
Bayesian Network Model	Koller and Friedman [64]
BLR08	Blum et al. [65]
Budget Absorption (BA), Budget Distribution (BD)	Kellaris et al. [18]
CFDS12	Chen et al. [66]
Data Recycling	Xiao et al. [67]
Dirichlet Prior	Machanavajjhala et al. [68]
Distributed Euler Histograms (DEH)	Xie et al. [69]
DJW13	Duchi et al. [70]
Edge Removal	Blocki et al. [71]
Eigen-Design	Li and Miklau [72]
FlowGraph	Raskhodnikova and Smith [73]
Fourier Perturbation Algorithm (FPA)	Barak et al. [74]
FreqItem	Zeng et al. [75]
Geometric Mechanism	Ghosh et al. [76]
Grouping and Smoothing (GS)	Kellaris and Papadopoulos [77]
Matrix Mechanism (MM)	Li et al. [78]
MWEM	Hardt et al. [79]
n-grams	Chen et al. [80]
Private Multiplicative Weights (PMW)	Hardt and Rothblum [81]
Private Record Matching	Inan et al. [82]
Private Spatial Decompositions (PSD)	Cormode et al. [83]
RAPPOR	Erlingsson et al. [84]
Sampling Perturbation Algorithm (SPA)	Rastogi and Nath [85]
Truncation	Kasiviswanathan et al. [86]
V-opt hist	Jagadish et al. [87]

Table 6: Ledger for papers outside of the SLR.

Algorithm	Internal Comparison	External Comparison
Boost	-	-
PMost, BMax	-	-
Privelet, Privelet^{+}, Privelet*	-	-
EFPA, P-HP	Boost, Privelet,NF, SF	SPA [85], MWEM [79]
NF, SF	Boost, Privelet	-
DPCopula	Privelet^{+}, P-HP	FP [88], PSD [83]
CiTM	-	-
PeGS, PeGS.rs	-	-
DPCube	Boost	Private Interactive ID3 [89], Private Record Matching [82]
PrivBayes	-	FPA[74], PrivGene [90], ERM [91]
AHP	NF, SF, P-HP	GS [77]
RG	-	BA [18], FAST [92]
ADMM	Boost, EFPA, P-HP, Privelet	LMM [78], RM [93]
DSAT, DSFT	-	-
(θ,Ω)-Histogram, θ-CumHisto	-	EdgeRemoval [71], Truncation [86], FlowGraph [73]
BPM	-	EM [13], Binary RR [70, 84]
PrivTree	Privelet*	UG [94–96], AG [94], Hierarchy [95], DAWA [97]
DPCocGen	PrivBayes	-
SORTaki	-	-
Pythia, Delphi	-	-
Tru, Min, Opt	Boost	n-grams [80], FreqItem [75], GS, DAWA, DPT [98]
DPPro	-	Private SVM [99], PriView [100], JTree [101]
T$^{\lambda}$	-	-
GGA	DSAT	-
PriSH	-	MWEM, DAWA
IHP, mIHP	Boost,EFPA, P-HP, SF, AHP	PSD, GS
RCF	Boost, NF	SHP [102]

Table 7: Algorithms used in empirical comparisons, divided by internal (included in the SLR) and external (excluded from the SLR) algorithms, sorted by year of publication. Comparisons with the Laplace mechanism and the author's own defined baselines (such as optimal) have been excluded from the table.

5.2 Categorization of Differentially Private Accuracy Improving Techniques

We observe from the algorithms in the 27 papers, there are three different dimensions to accuracy improvement in the context of differential privacy: **i)** *total noise reduction*, **ii)** *sensitivity reduction* and **iii)** *dimensionality reduction*.

i) **Total Noise Reduction** On the one hand, a histogram is published as statistical representation of a given data set (Goal I). On the other hand, histograms are published as a way to approximate the underlying distribution, which is then used to answer queries on the data set (Goal II). We refer to the latter as universal histograms: terminology adapted from [29]. In this dimension, optimizing the noisy end result (i.e differentially private histograms) provides opportunities for accuracy improvement.

ii) **Sensitivity Reduction** The global sensitivity of histogram queries is not small for graph data sets. Because, even a *relatively* small change in the network structure results in big change in the query answer. The accuracy improvement in this dimension follow from global sensitivity optimization.

iii) **Dimensionality Reduction** Publishing synthetic version of an entire data set consists of building a private statistical model from the original data set and then sampling data points from the model. In this dimension, inferring the underlying data distribution from a smaller set of attributes provides opportunities for accuracy improvement.

5.2.1 Dimension: Total Noise Reduction

In Table 8, we summarize the distinct techniques/approaches of the state-of-the-art from the point of view of reducing the total noise.

- ▷ **Goal I:** When the goal is to publish some statistical summary of a given data set as a differentially private histogram, histogram partitions play an essential role in improving the accuracy of the end result. A histogram partitioned into finer bins reduces approximation error[3] of the result, because each data point is correctly represented by a bin. However, the Laplace mechanism for histograms adds noise of scale $\Delta f/\varepsilon$ to each histogram bin. In other words, a histogram that is structured to minimize the approximation error, would suffer more noise in order to satisfy differential privacy.

The most common approach to enhance the utility for this goal, is to identify optimal histogram partitions for the given data.

Algorithms P-HP, SF and (θ,Ω)-Histogram use the Exponential mechanism to find V-optimal histogram [87] partitions. However, the quality of the partitions drops as the privacy budget available for iterating the Exponential mechanism decreases. Hence, algorithms NF, AHP, DPCocGen instead operate on the non-optimized noisy histogram for identifying suboptimal partitions for the final histogram. To further improve the quality of the partitions that are based on the non-optimized noisy histogram, in AHPsorting technique is used.

For the same goal described above, if the given data are bitmap strings then one opportunity for accuracy improvement is to vary the amount of noise for various histogram bins.

[3]Error caused by approximating the underlying distribution of data into histogram bins: intervals covering the range of domain values.

Algorithm BPM uses a bi-partite cut approach to partition a weighted histogram into bins with high average weight and bins with low relative weight. Further, in BPM the privacy budget ε is carefully split between the bins such that the heavy hitters, i.e. bins with high count, enjoy less noise. Algorithm AC uses weighted combination approach in terms of least square method in order to find optimal histogram partitions. Sample expansion through recycling the data points is another interesting approach for enhancing the accuracy of histograms over bitmap strings.

In the case of dynamic data sets, it is desirable to sequentially release the statistical summary of evolving data set at a given point in time. The most common approach is to limit the release of histograms, when there is a change in the data set for avoiding early depletion of privacy budget. Algorithms DSFT, DSAT and GGA uses distance-based sampling to monitor significant updates to the input data set. In algorithm RG an adaptive sampling process uses Bernoulli sampling for change detection in the data set. Further, in RG a novel histogram partitioning approach called retroactive grouping is introduced to enhance the accuracy of the end result.

- ▷ **Goal II:** When the histograms are used to answer workload of allowable queries. Laplace noise accumulates (sequential composition) as the number of queried histogram bins increases in order to answer the workload (covering large ranges of domain values). However, if the answer to the workload can be constructed by finding a linear combination of fewer bins, then the accuracy of the final answer will be significantly improved.

Algorithms Boost, DPCube, PrivTree, C*i*TM and mIHP employ an approach, where the domain ranges are hierarchically structured, typically in a tree structure. The intuition is, to find the fewest number of internal nodes such that the union of these ranges equals the desired range in the workload. To further improve the accuracy in the context of sequential composition, algorithm C*i*TM uses composition rule-based privacy budget optimization. Transformation techniques such as wavelet transform (Privelet) and Fourier transform (EFPA) are also used to model linear combination of domain ranges.

Another approach to reduce the accumulate noise in the context of universal histograms is to contain the total noise below a threshold. In BMax the maximum noise variance of the end result is contained within a threshold.

Furthermore, constraints are imposed in the output space of possible answers, which are then verified in the post-processing step to identify more accurate answers in the output space.

Preserving the dependency constraint is important for answering range queries over spatial histograms. To this end, in algorithm PriSH, true distribution of the underlying data set is learned from private answers to carefully chosen informative queries. Separately, to estimate the tail distribution of the final noisy histogram, algorithm θ-CumHisto uses some prior distribution to reallocate count values.

Category	Technique/Approach	Algorithms	Notes
Clustering	Bi-partite	BPM	
	Bisection	P-HP	
	Bisection	IHP, mIHP	
	MODL co-clustering [103]	DPCocGen	
	Matrix decomposition	Privelet$^+$	
	Weighted combination	AC	Least Square Method
	Retroactive Grouping	RG	Thresholded
	Selecting Top k	EFPA	
		CiTM	Key/foreign-key Relationships
		Min	Query Overlap
		SF	V-optimality
		NF	
		AHP	V-optimality
		(θ,Ω)-Histogram	V-optimality
		T$^\lambda$	Equi-width
Consistency Check	Frequency Calibration	θ-CumHisto	Monotonicity Property
	Hierarchical Consistency	Opt	
	Least Square Minimization	Boost	
	Least Square Minimization	DPCube	
	Realizable model	CiTM	Linear-time Approximation
		PMost	Least Norm Problem
Hierarchical Decomposition	Binary Tree	Boost	
	kd-tree	DPCube	V-optimality
	Quadtree	PrivTree	
	Query Tree	CiTM	Correlation of i-Table Model
	Sequential Partitions	mIHP	t-value
Learning True Distribution	Reallocate Values	θ-CumHisto	Linear Regression, Powerlaw & Uniform distributions
	Rescaling Weights	PriSH	Query Absolute Error, Dependency Constraints
Privacy Budget Optimization	Composition rule-based	CiTM	
	Threshold-driven Release	DSAT, DSFT	Adaptive-distance Qualifier, Fixed-distance Qualifier
	Threshold-driven Release	GGA	Fixed-distance Qualifier
	Weighted	BPM	
Sampling	Bernoulli Sampling	RG	
	Data Recycling	DRPP	
Sorting		AHP	
Transformation	Wavelet Transform	Privelet	
	Fourier Transformation	EFPA	
Threshold	Qualifying Weight	PMost	
	Qualifying Source-of-noise	BMax	
	Qualifying Source-of-noise	Tru	
	Sanitization	AHP	
	Wavelet Thresholding	Privelet*	

Table 8: Categorization of techniques/approaches used by each algorithm for total noise reduction. Additional qualifiers of each techniques are captured as notes.

5.2.2 Dimension: Sensitivity Reduction

In Table 9, we summarize the distinct techniques/approaches of the state-of-the-art from the point of view of reducing the global sensitivity.

Category	Technique/Approach	Algorithms	Notes
Neighbor Relation	Redefine	C*i*TM	Propagation Constraints
Projection	Edge Addition	(θ,Ω)-Histogram θ-CumHisto	Network Degree Bounded
	Edge Deletion	T$^{\lambda}$	Mutual Connections Bounded

Table 9: Categorization of techniques/approaches used by each algorithms for sensitivity reduction. Additional qualifiers of each techniques are captured as notes.

In graph data sets, global sensitivity becomes unbounded, for example, change in a node and its edges, in the worst case affects the whole structure (i.e involving all the nodes) of the network under *node differential privacy*. Bounding the network degree is one of the common approaches for containing the global sensitivity for analysis under *node differential privacy*. Techniques, edge addition ((θ,Ω)-Histogram, θ-CumHisto) and edge deletion (T$^{\lambda}$) are used to bound the size of the graph. Consequently, the noise required to satisfy *node differential privacy* will be reduced.

When there exists no *standard* neighborhood definition for the differential privacy guarantee in the light of correlated data structures. In the C*i*TM algorithm that operates on relational databases with multiple relation correlations, the neighbor relation is redefined.

5.2.3 Dimension: Dimensionality Reduction

In Table 10, we summarize the distinct techniques/approaches of the state-of-the-art from the point of view of reducing the data dimensions.

The most common approach to accuracy improvement in this dimension is to build statistical models that approximate the full dimensional distribution of the data set from multiple set marginal distributions. Some of techniques to approximate joint distribution of a data set are Bayesian Network (PrivBayes) and Copula functions (DPCopula). Furthermore, projection techniques from high-dimensional space to low-dimensional sub-spaces are shown to improve accuracy as less noise is required to make the smaller set of low-dimensional sub-spaces differentially private. Projection techniques found in the literature are, feature hashing using the hashing trick (PeGS) and random projection based on the Johnson-Lindenstrauss Lemma (DPPro).

In DPCopula, eigenvalue procedure is used in the post-processing stage to achieve additional gain in accuracy. Unexpectedly, reset-then-sample approach grouped under privacy budget optimization algorithmic category appear in this dimension, because the PeGS.rs algorithm supports multiple synthetic data set instances.

Category	Technique/Approach	Algorithms	Notes
Consistency check	Eigenvalue Procedure [104]	DPCopula	
Projection	Hashing Trick [105]	PeGS	
Privacy Budget Optimization	Reset-then-sample	PeGS.rs	
Transformation	Bayesian Network	PrivBayes	
	Copula Functions	DPCopula	
	Random Projection	DPPro	Johnson-Lindenstrauss Lemma

Table 10: Categorization of techniques/approaches used by each algorithm for data dimensionality reduction. Additional qualifiers of each techniques are captured as notes.

5.2.4 Summary

Figure 4 summarizes the categorization of differentially private accuracy improving techniques. Techniques identified in each accuracy improving dimensions are grouped into specific categories. The algorithmic categories are further partially sub-divided by the input data they support. Query answer relates to the type of release rather than to the input data, but the assumption is that the other mentioned data types, they implicitly specify the type of release.

The further the algorithmic category is located from the center of the circle, the more common is that category in that particular accuracy improvement dimension. Subsequently, clustering is the most commonly employed category for the total noise reduction dimension. Interestingly, same set of categories of accuracy improving algorithms are employed for dynamic data and bitmap strings, in the context of total noise reduction dimension. Hierarchical decomposition, consistency check and learning true distribution are primarily used in the context of releasing a histogram for answering workload of queries. It should be noted that the consistency check technique is used in the dimensionality reduction dimension as well but the usage of the technique is conditional.

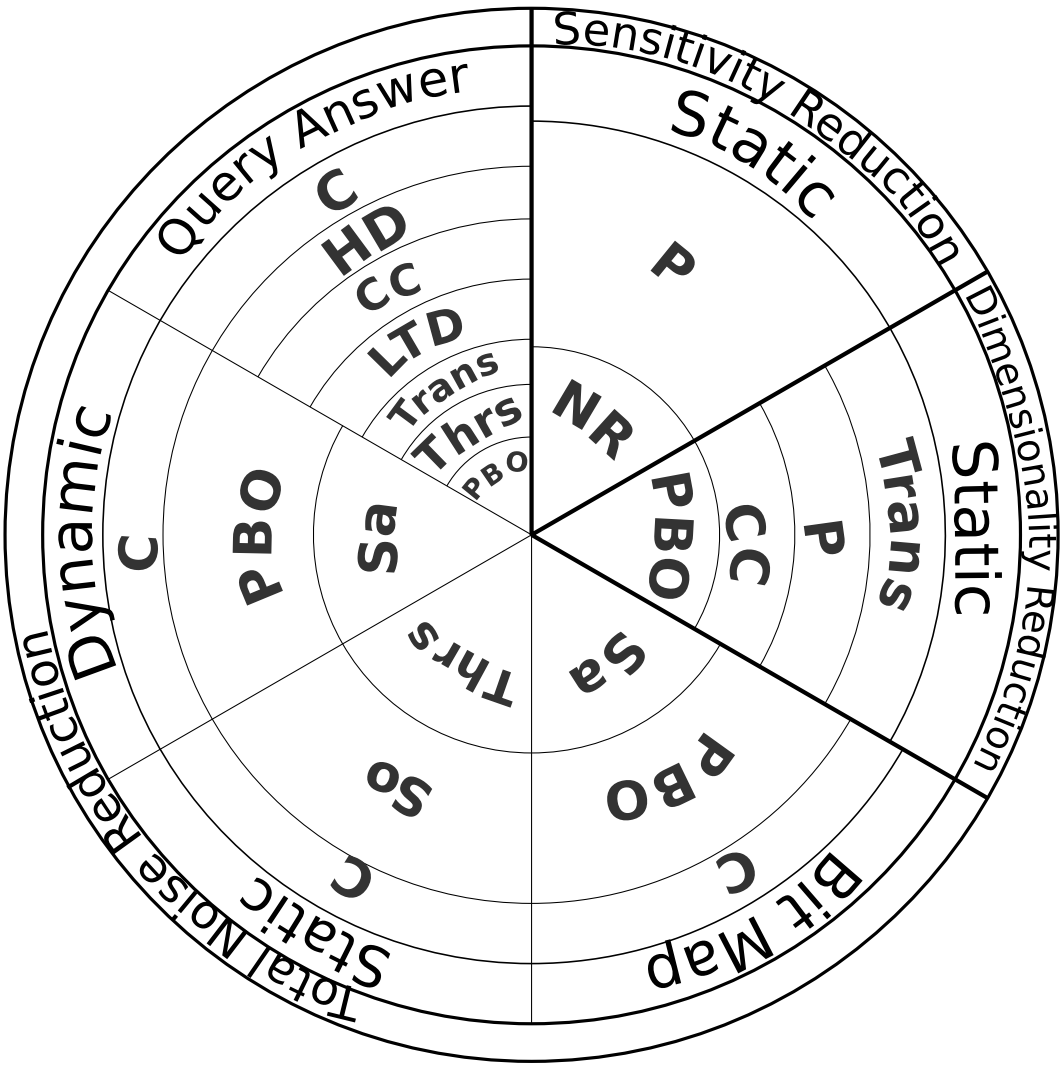

Figure 4: Conceptualization of accuracy improving techniques in the context of differential privacy: Abbreviations: **C**: Clustering, **CC**: Consistency Check, **HD**: Hierarchical Decomposition, **LTD**: Learning True Distribution, **NR**: Neighborhood Redefine, **P**: Projection, **PBO**: Privacy Budget Optimization, **Thrs**: Threshold, **Trans**: Transformation, **Sa**: Sampling, **So**: Sorting.

6 Discussion and Open Challenges

One limitation of this paper is that the scope of our SLR is limited to papers with empirical results. We have chosen empirical measurement of accuracy, since it can provide a less pessimistic understanding of error bounds, as opposed to analytical bounds. However, in our analysis (Section 5) of the papers, we studied related theoretical aspects of accuracy improvements and put the surveyed papers into context by tracing their origin, illustrated in Figure 3. As such, we can guide the interested reader in the right direction, but we do not provide an analysis of theoretical results.

Next (Section 6.1), we identify possible future work, mainly related to composability of the different techniques. It is not clear exactly which techniques compose, or how many techniques from each place that can be used to achieve accuracy improvements. Hence, open challenges include both coming up with new accuracy techniques for each place as well as combining techniques in meaningful, composable ways. Last Section 6.2, we list the papers that were excluded as part of our qualitative analysis.

6.1 Composability of Categories

From the dimensions identified in our analysis, we continue by investigating how techniques from different categories *may* be composed. We also connect the papers with the *place* [4] their algorithm operates on in Table 11.

We believe a technique from one place is possible to compose with techniques from another place, since the places are designed to be a sequential representation of the data analysis. An open challenge derived from Table 11 is boosting each algorithm's accuracy by adding more techniques, either in a place which does not yet have any accuracy improvement, or together with the already existing techniques. For example, an algorithm that has improvement in place B (post-processing) may be combined with place A, C and/or D. Similarly, it may be possible to compose one technique from place B with another technique also from place B.

Next, we will illustrate how composability is already achieved by giving a few examples of how techniques are composed in the included papers.

Place A: Altering the Query

Altering the query targets *sensitivity reduction*, as sensitivity is a property of the query. Our take away from the SLR is that there are mainly two tricks to altering the query:

1. When an analysis requires a high sensitivity query, replace the query with an approximate query, or break the query down into two or more sub-queries.

2. Use sampling to avoid prematurely exhausting the privacy budget.

Item 1: For example, a histogram query is broken down into two separate queries: a clustering technique based on the exponential mechanism and usually a Laplace counting query, as in the case with Xu et al. [33] and consecutive work.

By breaking down the query, the sensitivity reduction can increase accuracy, but it needs to be balanced against the source of accumulated noise that is introduced by multiple queries. In particular, when breaking down a query, the privacy budget needs to be appropriately

[4] Places refers to different points in the workflow of a typical differentially private analysis, see Figure 1

	A	B	C	D
Hay et al. [29]		✓		
Ding et al. [30]		✓	✓	
Xiao et al. [31]				✓
Ács et al. [32]			✓	✓
Xu et al. [33]			✓	
Li et al. [34]			✓	
Lu et al. [35]		✓		✓
Park and Ghosh [36]		✓		✓
Xiao et al. [37]			✓	
Zhang et al. [38]			✓	
Zhang et al. [39]			✓	
Chen et al. [40]		✓		
Lee et al. [41]		✓		
Li et al. [42]	✓		✓	
Day et al. [43]	✓	✓	✓	
Wang et al. [44]			✓	
Zhang et al. [45]			✓	
Benkhelif et al. [46]		✓	✓	
Doudalis and Mehrotra [47]			✓	
Kotsogiannis et al. [48]	✓	✓		✓
Wang et al. [49]			✓	
Xu et al. [50]			✓	
Ding et al. [51]	✓			✓
Gao and Ma [52]			✓	
Ghane et al. [53]		✓	✓	
Li et al. [54]			✓	
Nie et al. [55]		✓		✓

Table 11: Mapping the papers to each place where: A) Altering the query, B) Post-processing, C) Change in mechanism, D) Pre-processing.

distributed between the sub-queries. For example, when breaking a histogram into a clustering query and then a count query, one could choose to give more budget to the clustering step to find a tighter histogram structure, but that would come at the cost of less accuracy for the count query.

Item 2: When an analysis is done on dynamic data, it is possible to unintentionally include the same data points in multiple queries, and ending up 'paying' for them multiple times. Li et al. [42] mitigates this source of accumulated noise by deploying sampling. It is also possible to use sampling for static data, for example, Delphi by Kotsogiannis et al. [48] could be trained on a sample of the full data set, if no public training data is available.

Place B: Post-processing

Post-processing targets *total noise reduction*, usually by exploiting consistency checks or other known constraints. Since post-processing is done on data that has been released by a differentially private algorithm, post-processing can always be done without increasing the privacy loss. However, post-processing can still decrease accuracy if used carelessly. In our SLR, the main post-processing idea is:

1. Finding approximate solutions to get rid of inconsistencies through *constrained inference* [29].

2. Applying consistency checks that would hold for the raw data.

Item 1: Boost is already being combined with several algorithms that release histograms, for example NF and SF. ADMM is a similar, but more generic solution that has been applied to more output types than just histograms. In fact, Lee et al. [41] claims ADMM can re-use algorithms use for least square minimization, which means Boost should be possible to incorporate in ADMM. Consequently, we believe ADMM would compose with most algorithms due to its generic nature.

Place C: Change in the Release Mechanism

Changing the release mechanism mainly targets *total noise reduction*. In the SLR, we found the following approaches being used:

1. Test-and-release.
2. Sorting as an intermediary step.

Item 1: DSAT and DSFT uses thresholding to determine when to release data, as a way to save the privacy budget. Thresholding is particularly useful for dynamic data, as it often requires multiple releases over time. For example, adaptive or fixed thresholding can be used for sensor data and trajectory data, effectively providing a way of sampling the data.
SF also uses a type of test-and-release when creating the histogram structure using the exponential mechanism. The test-and-release approach means EM can be combined with basically any other release mechanism, which is also what we found in the literature. We believe the main challenge with EM is finding an adequate scoring/utility function, and this is where we believe a lot of accuracy improvement will come from.
Item 2 SORTaki is designed to be composable with two-step algorithms that release histograms, for example NF. The idea is that by sorting noisy values, they can group together similar values that would otherwise not be grouped due to the bins not being adjacent.

Place D: Pre-processing

Pre-processing generally targets *dimensionality reduction* or *total noise reduction*. In our SLR, we encountered the following types of pre-processing:

1. Encoding through projection/transformation.
2. Learning on non-sensitive data.

Item 1: Several algorithms project or transform their data, for example Privelet and EFPA. Encoding can reduce both sensitivity and dimensionality by decreasing redundancy, and is therefore especially interesting for multi-dimensional as well as high-dimensional, sparse, data sets. However, lossy compression techniques can potentially introduce new sources of noise, and therefore adds another trade-off that needs to be taken into account. Intuitively, lossy compression is beneficial when the noise lost in the compression step is greater than the proportion of useful data points lost. For example, sparse data may benefit more from lossy compression than data that is not sparse.
Item 2: Delphi is a pre-processing step which uses a non-sensitive, public data set to build a decision tree. In cases where public data sets are available, it could be possible to adopt the same idea; for example learning a histogram structure on public data as opposed to spending budget on it. The caveat here is of course that the public data needs to be similar enough

to the data used in the differentially private analysis, because otherwise this becomes an added source of noise. Thus, learning from non-sensitive data introduces another trade-off that is still largely unexplored.

6.2 Incomparable papers

We present a list of papers that were excluded during our qualitative analysis, and the reason for why we decided to exclude them in Section 5. The reason for excluding papers in the analysis step is that certain properties of their algorithms make them incomparable with other algorithms.

[106]: The DP-FC algorithm does not consider the structure of a histogram a sensitive attribute, and thus achieves a trivial accuracy improvement over other algorithms.

[107]: The APG algorithm does not perform differentially private clustering, and therefore achieves better accuracy by relaxing the privacy guarantees compared to AHP, IHP and GS.

[108]: The SC algorithm uses the ordering of the bins in order to calculate the cluster centers, but does not perturb the values before doing so, and thus the order is not protected, making their guarantees incomparable.

[109]: The Outlier-Histopub algorithm, similarly sorts the bin counts according to size, without using the privacy budget accordingly to learn this information. The authors claim that this type of sorting does not violate differential privacy, but due to the fact that the output is determined based on the private data, the approach cannot be 0-differentially private.

[110]: The ASDP-HPA algorithm does not describe the details of how their use of Autoregressive Integrated Moving Average Model (ARIMA) is made private, and thus we cannot determine whether the entire algorithm is differentially private. Furthermore, the details of how they pre-process their data set is not divulged, and it can thus not be determined if the pre-processing violates differential privacy or not by changing the query sensitivity.

[111]: The algorithm is incomplete, since it only covers the histogram partitioning, and does not involve the addition of noise to bins. Furthermore, it is not clear whether they draw noise twice using the same budget, or if they reuse the same noise for their thresholds. As the privacy guarantee ε cannot be unambiguously deduced, we do not include their paper in our comparison.

[112]: The GBLUE algorithm generates a k-range tree based on the private data, where k is the fanout of the tree. Since private data is used to decide on whether a node is further split or not, it does not provide the same privacy guarantees as the other studied algorithms.

[113]: The algorithm creates groups based on the condition that the merged bins guarantee k-indistinguishability. Since this merge condition is based on the property of the data it does not guarantee differential privacy on the same level as the other papers, so we deem it incomparable.

Further, in the analysis regarding dimensions of accuracy improvement techniques presented in Section 5, some algorithms such as ADMM, SORTaki and Pythia are excluded. The rationale behind the exclusion is, these algorithms are not self contained, but nevertheless improves accuracy of the differentially private answers when combined with other analyzed algorithms.

Efforts such as Pythia and DPBench [114], that provide practitioners a way to empirically assess the privacy/accuracy trade-off related to their data sets are commendable. However, to effectively use the tool one needs to have some background knowledge of the right combination of parameters to tune. In our analysis of the algorithms, we mapped out the accuracy improvement techniques grouped by optimization goals and corresponding query size. This knowledge will allow practitioners and researchers alike to think about other places to explore for accuracy improvement, rather than finding the algorithms that are based only on their data. Essentially, we provide an understanding to enable algorithm design, as opposed to algorithm selection.

7 Conclusions

Motivated by scarcity of works that structure knowledge concerning accuracy improvement in differentially private computations, we conducted a systematic literature review (SLR) on accuracy improvement techniques for histogram and synthetic data publication under differential privacy.

We present two results from our analysis that addresses our research objective, namely to synthesize the understanding of the underlying foundations of the privacy/accuracy trade-off in differentially private computations. This systematization of knowledge (SoK) includes:

1. Internal/external positioning of the studied algorithms (Figure 3 and Table 7).

2. A taxonomy of different *categories* (Figure 4) and their corresponding *optimization goals* to achieve accuracy improvement: *total noise reduction* (Table 8), *sensitivity reduction* (Table 9) and *data dimensionality reduction* (Table 10).

What's more, we also discuss and present an overview of composable algorithms according to their optimization goals and category, sort-out by the *places*, in which they operate (Section 6.1). Our intent is that these findings will pave the way for future research by allowing others to integrate new solutions according to the categories. For example, our places can be used to reason about where to plug in new or existing techniques targeting a desired *optimization goal* during algorithm design.

From our overview of composability, we see that most efforts are focused on making *changes in the mechanism*, and on *post-processing*. We observe that, *altering the query* in one way or another, is not popular, and we believe further investigation is required to understand which techniques can be adopted in this place.

Finally, although all algorithms focus on accuracy improvement, it is impossible to select the 'best' algorithm without context. Intuitively, newer algorithms will have improved some property of an older algorithm, meaning that newer algorithms *may* provide higher accuracy. Still, the algorithms are used for different analyses, which means not all algorithms will be interchangeable. Secondly, many algorithms are data dependent, which means that the selection of the 'best' algorithm may change depending on the input data

used, even when the analysis is fixed. Consequently, the 'best' algorithm needs to be chosen with a given data set and a given analysis in mind. The problem of choosing the 'best' algorithm when the setting is known is in fact addressed by Pythia.

Acknowledgements

Boel Nelson was partly funded by the Swedish Foundation for Strategic Research (SSF) and the Swedish Research Council (VR).

References

[1] P. Samarati and L. Sweeney. *Protecting Privacy When Disclosing Information: K-Anonymity and Its Enforcement through Generalization and Suppression*. Technical Report. SRI International, 1998.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. "L-Diversity: Privacy beyond k -Anonymity". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007).

[3] N. Li, T. Li, and S. Venkatasubramanian. "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2007.

[4] A. Narayanan and V. Shmatikov. "Myths and Fallacies of "Personally Identifiable Information"". In: *Communications of the ACM* 53.6 (2010), pp. 24–26.

[5] C. Dwork. "Differential Privacy". In: *International Colloquium on Automata, Languages and Programming (ICALP)*. Springer, 2006.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography Conference (TCC)*. Springer, 2006.

[7] C. M. Bowen and F. Liu. "Comparative Study of Differentially Private Data Synthesis Methods". arXiv preprint arXiv:1911.12704, 2016.

[8] H. Li, L. Xiong, and X. Jiang. "Differentially Private Histogram and Synthetic Data Publication". In: *Medical Data Privacy Handbook*. Ed. by A. Gkoulalas-Divanis and G. Loukides. Cham: Springer, 2015, pp. 35–58.

[9] X. Meng, H. Li, and J. Cui. "Different Strategies for Differentially Private Histogram Publication". In: *Journal of Communications and Information Networks* 2.3 (2017), pp. 68–77.

[10] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. "Our Data, Ourselves: Privacy Via Distributed Noise Generation". In: *International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*. Ed. by S. Vaudenay. Springer, 2006.

[11] S. Meiser. "Approximate and Probabilistic Differential Privacy Definitions". In: *IACR Cryptology ePrint Archive* (2018), p. 9.

[12] C. Dwork and A. Roth. "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends in Theoretical Computer Science* 9.3-4 (2014), pp. 211–407.

[13] F. McSherry and K. Talwar. "Mechanism Design via Differential Privacy". In: *Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2007.

[14] C. Dwork. "Differential Privacy: A Survey of Results". In: *Theory and Applications of Models of Computation*. Ed. by M. Agrawal, D. Du, Z. Duan, and A. Li. LNCS 4978. Springer Berlin Heidelberg, 2008, pp. 1–19.

[15] M. Hay, C. Li, G. Miklau, and D. Jensen. "Accurate Estimation of the Degree Distribution of Private Networks". In: *International Conference on Data Mining (ICDM)*. IEEE, 2009.

[16] C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin. "Pan-Private Streaming Algorithms". In: *Innovations in Computer Science (ICS)*. Tsinghua University Press, 2010.

[17] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum. "Differential Privacy under Continual Observation". In: *Symposium on Theory of Computing (STOC)*. ACM, 2010.

[18] G. Kellaris, S. Papadopoulos, X. Xiao, and D. Papadias. "Differentially Private Event Sequences over Infinite Streams". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2014.

[19] B. Kitchenham. *Procedures for Performing Systematic Reviews*. Joint Technical Report TR/SE-0401. Keele University, 2004.

[20] B. A. Kitchenham, T. Dyba, and M. Jorgensen. "Evidence-Based Software Engineering". In: *Proceedings. 26th International Conference on Software Engineering*. IEEE, 2004, pp. 273–281.

[21] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang. "An Overview of Microsoft Academic Service (MAS) and Applications". In: *International Conference on World Wide Web (WWW)*. ACM, 2015.

[22] Microsoft. *Microsoft Academic*. 2019. URL: `https://academic.microsoft.com/home` (visited on 10/09/2019).

[23] A.-W. Harzing. "Microsoft Academic (Search): A Phoenix Arisen from the Ashes?" en. In: *Scientometrics* 108.3 (2016), pp. 1637–1647.

[24] A.-W. Harzing and S. Alakangas. "Microsoft Academic Is One Year Old: The Phoenix Is Ready to Leave the Nest". en. In: *Scientometrics* 112.3 (2017), pp. 1887–1894.

[25] A.-W. Harzing and S. Alakangas. "Microsoft Academic: Is the Phoenix Getting Wings?" en. In: *Scientometrics* 110.1 (2017), pp. 371–383.

[26] S. E. Hug and M. P. Brändle. "The Coverage of Microsoft Academic: Analyzing the Publication Output of a University". en. In: *Scientometrics* 113.3 (2017), pp. 1551–1571.

[27] A.-W. Harzing. "Two New Kids on the Block: How Do Crossref and Dimensions Compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science?" en. In: *Scientometrics* 120.1 (2019), pp. 341–349.

[28] R. C. Nickerson, U. Varshney, and J. Muntermann. "A Method for Taxonomy Development and Its Application in Information Systems". In: *European Journal of Information Systems* 22.3 (2013), pp. 336–359.

[29] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. "Boosting the Accuracy of Differentially Private Histograms through Consistency". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2010.

[30] B. Ding, M. Winslett, J. Han, and Z. Li. "Differentially Private Data Cubes: Optimizing Noise Sources and Consistency". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2011.

[31] X. Xiao, G. Wang, and J. Gehrke. "Differential Privacy via Wavelet Transforms". In: *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 23.8 (2011), pp. 1200–1214.

[32] G. Ács, C. Castelluccia, and R. Chen. "Differentially Private Histogram Publishing through Lossy Compression". In: *International Conference on Data Mining (ICDM)*. IEEE, 2012.

[33] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett. "Differentially Private Histogram Publication". In: *The VLDB Journal* 22.6 (2013), pp. 797–822.

[34] H. Li, L. Xiong, and X. Jiang. "Differentially Private Synthesization of Multi-Dimensional Data Using Copula Functions". In: *International Conference on Extending Database Technology (EDBT)*. Vol. 2014. NIH Public Access, 2014.

[35] W. Lu, G. Miklau, and V. Gupta. "Generating Private Synthetic Databases for Untrusted System Evaluation". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2014.

[36] Y. Park and J. Ghosh. "PeGS: Perturbed Gibbs Samplers That Generate Privacy-Compliant Synthetic Data". In: *Transactions on Data Privacy (TDP)* 7.3 (2014), pp. 253–282.

[37] Y. Xiao, L. Xiong, L. Fan, S. Goryczka, and H. Li. "DPCube: Differentially Private Histogram Release through Multidimensional Partitioning". In: *Transactions on Data Privacy (TDP)* 7.3 (2014), pp. 195–222.

[38] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. "PrivBayes: Private Data Release via Bayesian Networks". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2014.

[39] X. Zhang, R. Chen, J. Xu, X. Meng, and Y. Xie. "Towards Accurate Histogram Publication under Differential Privacy". In: *International Conference on Data Mining (SDM)*. SIAM, 2014.

[40] R. Chen, Y. Shen, and H. Jin. "Private Analysis of Infinite Data Streams via Retroactive Grouping". In: *International on Conference on Information and Knowledge Management (CIKM)*. ACM, 2015.

[41] J. Lee, Y. Wang, and D. Kifer. "Maximum Likelihood Postprocessing for Differential Privacy under Consistency Constraints". In: *International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2015.

[42] H. Li, L. Xiong, X. Jiang, and J. Liu. "Differentially Private Histogram Publication for Dynamic Datasets: An Adaptive Sampling Approach". In: *International on Conference on Information and Knowledge Management (CIKM)*. ACM, 2015.

[43] W.-Y. Day, N. Li, and M. Lyu. "Publishing Graph Degree Distribution with Node Differential Privacy". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2016.

[44] S. Wang, L. Huang, P. Wang, H. Deng, H. Xu, and W. Yang. "Private Weighted Histogram Aggregation in Crowdsourcing". In: *International Conference on Wireless Algorithms, Systems, and Applications (WASA)*. Springer, 2016.

[45] J. Zhang, X. Xiao, and X. Xie. "PrivTree: A Differentially Private Algorithm for Hierarchical Decompositions". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2016.

[46] T. Benkhelif, F. Fessant, F. Clérot, and G. Raschia. "Co-Clustering for Differentially Private Synthetic Data Generation". In: *International Workshop on Personal Analytics and Privacy (PAP)*. Springer, 2017.

[47] S. Doudalis and S. Mehrotra. "SORTaki: A Framework to Integrate Sorting with Differential Private Histogramming Algorithms". In: *Conference on Privacy, Security and Trust (PST)*. IEEE, 2017.

[48] I. Kotsogiannis, A. Machanavajjhala, M. Hay, and G. Miklau. "Pythia: Data Dependent Differentially Private Algorithm Selection". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2017.

[49] N. Wang, Y. Gu, J. Xu, F.-F. Li, and G. Yu. "Differentially Private Event Histogram Publication on Sequences over Graphs". In: *Journal of Computer Science and Technology* 32.5 (2017), pp. 1008–1024.

[50] C. Xu, J. Ren, Y. Zhang, Z. Qin, and K. Ren. "DPPro: Differentially Private High-Dimensional Data Release via Random Projection". In: *IEEE Transactions on Information Forensics and Security* 12.12 (2017), pp. 3081–3093.

[51] X. Ding, X. Zhang, Z. Bao, and H. Jin. "Privacy-Preserving Triangle Counting in Large Graphs". In: *International Conference on Information and Knowledge Management (CIKM)*. ACM, 2018.

[52] R. Gao and X. Ma. "Dynamic Data Histogram Publishing Based on Differential Privacy". In: *International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/ BDCloud/SocialCom/SustainCom)*. IEEE, 2018.

[53] S. Ghane, L. Kulik, and K. Ramamohanarao. "Publishing Spatial Histograms under Differential Privacy". In: *International Conference on Scientific and Statistical Database Management (SSDBM)*. ACM, 2018.

[54] H. Li, J. Cui, X. Meng, and J. Ma. "IHP: Improving the Utility in Differential Private Histogram Publication". In: *Distributed and Parallel Databases* 37 (2019), pp. 721–750.

[55] Y. Nie, W. Yang, L. Huang, X. Xie, Z. Zhao, and S. Wang. "A Utility-Optimized Framework for Personalized Private Histogram Estimation". In: *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 31.4 (2019), pp. 655–669.

[56] X. Xiao, G. Wang, and J. Gehrke. "Differential Privacy via Wavelet Transforms". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2010.

[57] J. Xu, Z. Zhang, X. Xiao, Y. Yang, and G. Yu. "Differentially Private Histogram Publication". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2012.

[58] Y. Park, J. Ghosh, and M. Shankar. "Perturbed Gibbs Samplers for Generating Large-Scale Privacy-Safe Synthetic Health Data". In: *International Conference on Healthcare Informatics*. IEEE, 2013.

[59] Y. Xiao, L. i Xiong, and C. Yuan. "Differentially Private Data Release through Multidimensional Partitioning". In: *Workshop on Secure Data Management (SDM)*. Springer, 2010.

[60] Y. Xiao, J. Gardner, and L. Xiong. "DPCube: Releasing Differentially Private Data Cubes for Health Information". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2012.

[61] H. Li, J. Cui, X. Lin, and J. Ma. "Improving the Utility in Differential Private Histogram Publishing: Theoretical Study and Practice". In: *International Conference on Big Data (IEEE Big Data)*. IEEE, 2016.

[62] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". In: *Foundations and Trends® in Machine Learning* 3.1 (2011), pp. 1–122.

[63] A. Arasu, R. Kaushik, and J. Li. "Data Generation Using Declarative Constraints". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2011.

[64] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[65] A. Blum, K. Ligett, and A. Roth. "A Learning Theory Approach to Non-Interactive Database Privacy". In: *Symposium on Theory of Computing (STOC)*. ACM, 2008.

[66] R. Chen, B. C. Fung, B. C. Desai, and N. M. Sossou. "Differentially Private Transit Data Publication: A Case Study on the Montreal Transportation System". In: *International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2012.

[67] X. Xiao, Y. Tao, and M. Chen. "Optimal Random Perturbation at Multiple Privacy Levels". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2009.

[68] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. "Privacy: Theory Meets Practice on the Map". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2008.

[69] H. Xie, E. Tanin, and L. Kulik. "Distributed Histograms for Processing Aggregate Data from Moving Objects". In: *International Conference on Mobile Data Management (MDM)*. IEEE, 2007.

[70] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. "Local Privacy and Statistical Minimax Rates". In: *Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2013.

[71] J. Blocki, A. Blum, A. Datta, and O. Sheffet. "Differentially Private Data Analysis of Social Networks via Restricted Sensitivity". In: *Conference on Innovations in Theoretical Computer Science (ITCS)*. ACM, 2013.

[72] C. Li and G. Miklau. "An Adaptive Mechanism for Accurate Query Answering Under Differential Privacy". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2012.

[73] S. Raskhodnikova and A. Smith. "Efficient Lipschitz Extensions for High-Dimensional Graph Statistics and Node Private Degree Distributions". arXiv preprint arXiv:1504.07912, 2015.

[74] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. "Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release". In: *Symposium on Principles of Database Systems (PODS)*. ACM, 2007.

[75] C. Zeng, J. F. Naughton, and J.-Y. Cai. "On Differentially Private Frequent Itemset Mining". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2012.

[76] A. Ghosh, T. Roughgarden, and M. Sundararajan. "Universally Utility-Maximizing Privacy Mechanisms". In: *SIAM Journal on Computing* 41.6 (2012), pp. 1673–1693.

[77] G. Kellaris and S. Papadopoulos. "Practical Differential Privacy via Grouping and Smoothing". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2013.

[78] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. "Optimizing Linear Counting Queries Under Differential Privacy". In: *Symposium on Principles of Database Systems (PODS)*. ACM, 2010.

[79] M. Hardt, K. Ligett, and F. Mcsherry. "A Simple and Practical Algorithm for Differentially Private Data Release". In: *Advances in Neural Information Processing Systems (NIPS)*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012.

[80] R. Chen, G. Acs, and C. Castelluccia. "Differentially Private Sequential Data Publication via Variable-Length N-Grams". In: *Conference on Computer and Communications Security (CCS)*. ACM, 2012.

[81] M. Hardt and G. N. Rothblum. "A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis". In: *Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2010.

[82] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino. "Private Record Matching Using Differential Privacy". In: *International Conference on Extending Database Technology (EDBT)*. ACM, 2010.

[83] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu. "Differentially Private Spatial Decompositions". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2012.

[84] U. Erlingsson, V. Pihur, and A. Korolova. "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response". In: *Conference on Computer and Communications Security (CCS)*. ACM, 2014.

[85] V. Rastogi and S. Nath. "Differentially Private Aggregation of Distributed Time-Series with Transformation and Encryption". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2010.

[86] S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith. "Analyzing Graphs with Node Differential Privacy". In: *Theory of Cryptography Conference (TCC)*. Ed. by A. Sahai. Springer, 2013.

[87] H. V. Jagadish, N. Koudas, S. Muthukrishnan, V. Poosalaa, K. Sevcik, and T. Suel. "Optimal Histograms with Quality Guarantees". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 1998.

[88] G. Cormode, C. Procopiuc, D. Srivastava, and T. T. L. Tran. "Differentially Private Summaries for Sparse Data". In: *International Conference on Database Theory (ICDT)*. ACM, 2012.

[89] A. Friedman and A. Schuster. "Data Mining with Differential Privacy". In: *International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2010.

[90] J. Zhang, X. Xiao, Y. Yang, Z. Zhang, and M. Winslett. "PrivGene: Differentially Private Model Fitting Using Genetic Algorithms". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2013.

[91] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. "Differentially Private Empirical Risk Minimization". In: *Journal of Machine Learning Research* 12 (Mar 2011), pp. 1069–1109.

[92] L. Fan and L. Xiong. "An Adaptive Approach to Real-Time Aggregate Monitoring With Differential Privacy". In: *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 26.9 (2014), pp. 2094–2106.

[93] G. Yuan, Z. Zhang, M. Winslett, X. Xiao, Y. Yang, and Z. Hao. "Low-Rank Mechanism: Optimizing Batch Queries under Differential Privacy". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2012.

[94] W. Qardaji, W. Yang, and N. Li. "Differentially Private Grids for Geospatial Data". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2013.

[95] W. Qardaji, W. Yang, and N. Li. "Understanding Hierarchical Methods for Differentially Private Histograms". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2013.

[96] D. Su, J. Cao, N. Li, E. Bertino, and H. Jin. "Differentially Private K-Means Clustering". In: *Conference on Data and Application Security and Privacy (CODASPY)*. ACM, 2016.

[97] C. Li, M. Hay, G. Miklau, and Y. Wang. "A Data- and Workload-Aware Algorithm for Range Queries Under Differential Privacy". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2014.

[98] X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, and D. Srivastava. "DPT: Differentially Private Trajectory Synthesis Using Hierarchical Reference Systems". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2015.

[99] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft. "Learning in a Large Function Space: Privacy-Preserving Mechanisms for SVM Learning". In: *Journal of Privacy and Confidentiality* 4.1 (2012).

[100] W. Qardaji, W. Yang, and N. Li. "PriView: Practical Differentially Private Release of Marginal Contingency Tables". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2014.

[101] R. Chen, Q. Xiao, Y. Zhang, and J. Xu. "Differentially Private High-Dimensional Data Publication via Sampling-Based Inference". In: *International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2015.

[102] R. Bassily and A. Smith. "Local, Private, Efficient Protocols for Succinct Histograms". In: *Symposium on Theory of Computing (STOC)*. ACM, 2015.

[103] M. Boullé. "Data Grid Models for Preparation and Modeling in Supervised Learning". In: *Hands-On Pattern Recognition: Challenges in Machine Learning* 1 (2011), pp. 99–130.

[104] P. J. Rousseeuw and G. Molenberghs. "Transformation of Non Positive Semidefinite Correlation Matrices". In: *Communications in Statistics - Theory and Methods* 22.4 (1993), pp. 965–984.

[105] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. "Feature Hashing for Large Scale Multitask Learning". In: *International Conference on Machine Learning (ICML)*. The 26th Annual International Conference. ACM, 2009.

[106] F. Yan, X. Zhang, C. Li, W. Li, S. Li, and F. Sun. "Differentially Private Histogram Publishing through Fractal Dimension for Dynamic Datasets". In: *Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2018.

[107] X. Liu and S. Li. "Histogram Publishing Method Based on Differential Privacy". In: *International Conference on Computer Science and Software Engineering (CSSE)* (2018).

[108] Q. Qian, Z. Li, P. Zhao, W. Chen, H. Yin, and L. Zhao. "Publishing Graph Node Strength Histogram with Edge Differential Privacy". In: *International Conference on Database Systems for Advanced Applications (DASFAA)*. Springer, 2018.

[109] Q. Han, B. Shao, L. Li, Z. Ma, H. Zhang, and X. Du. "Publishing Histograms with Outliers under Data Differential Privacy". In: *Security and Communication Networks* 9.14 (2016), pp. 2313–2322.

[110] Y. Li and S. Li. "Research on Differential Private Streaming Histogram Publication Algorithm". In: *International Conference on Cloud Computing and Intelligence Systems (CCIS)*. IEEE, 2018.

[111] M. Hadian, X. Liang, T. Altuwaiyan, and M. M. E. A. Mahmoud. "Privacy-Preserving mHealth Data Release with Pattern Consistency". In: *Global Communications Conference (GLOBECOM)*. IEEE, 2016.

[112] H. Chen, Y. Wu, T. Chen, and X. Wang. "An Iterative Algorithm for Differentially Private Histogram Publication". In: *International Conference on Cloud Computing and Big Data (CLOUDCOM-ASIA)*. IEEE, 2013.

[113] X. Li, J. Yang, Z. Sun, and J. Zhang. "Differential Privacy for Edge Weights in Social Networks". In: *Security and Communication Networks* 2017 (2017), pp. 1–10.

[114] M. Hay, A. Machanavajjhala, G. Miklau, Y. Chen, and D. Zhang. "Principled Evaluation of Differentially Private Algorithms Using DPBench". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2016.

[115] V. Balcer and S. P. Vadhan. "Differential Privacy on Finite Computers". In: *Conference on Innovations in Theoretical Computer Science (ITCS)*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018.

[116] T. Benkhelif. "Publication de Données Individuelles Respectueuse de La Vie Privée : Une Démarche Fondée Sur Le Co-Clustering". PhD thesis. Université de Nantes, 2018.

[117] A. Bhowmick, A. H. Vyrros, M. R. Salesi, and U. S. Vaishampayan. "Differential Privacy Using a Multibit Histogram". U.S. pat. 20180349620A1. Apple Inc. 2018.

[118] C. M. Bowen and F. Liu. "Differentially Private Release and Analysis of Youth Voter Registration Data via Statistical Election to Partition Sequentially". arXiv preprint arXiv:1602.01063, 2018.

[119] C. M. Bowen and F. Liu. "STatistical Election to Partition Sequentially (STEPS) and Its Application in Differentially Private Release and Analysis of Youth Voter Registration Data". arXiv preprint arXiv:1803.06763, 2018.

[120] K. Chaudhuri and S. A. Vinterbo. "A Stability-Based Validation Procedure for Differentially Private Machine Learning". In: *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2013.

[121] B. Cyphers and K. Veeramachaneni. "AnonML: Locally Private Machine Learning over a Network of Peers". In: *International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2017.

[122] E. C. Eugenio and F. Liu. "CIPHER: Construction of dIfferentially Private Microdata from Low-Dimensional Histograms via Solving Linear Equations with Tikhonov Regularization." arXiv preprint arXiv:1812.05671, 2018.

[123] M. Fanaeepour, L. Kulik, E. Tanin, and B. I. P. Rubinstein. "The CASE Histogram: Privacy-Aware Processing of Trajectory Data Using Aggregates". In: *Geoinformatica* 19.4 (2015), pp. 747–798.

[124] M. Fanaeepour and B. I. P. Rubinstein. "End-to-End Differentially-Private Parameter Tuning in Spatial Histograms." arXiv preprint arXiv:1702.05607, 2017.

[125] M. Fanaeepour and B. I. P. Rubinstein. "Histogramming Privately Ever After: Differentially-Private Data-Dependent Error Bound Optimisation". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2018.

[126] Y. Fei, Z. Xing, L. Chang, S. Wei, L. Wanjie, and L. Shuai. "Differential Privacy Protection-Based Data Release Method for Spark Framework". Pat. CN107766740A (China). 2018.

[127] A. Foote, A. Machanavajjhala, and K. McKinney. *Releasing Earnings Distributions Using Differential Privacy: Disclosure Avoidance System For Post Secondary Employment Outcomes (PSEO)*. Economic Analysis. 2019.

[128] J. J. Gardner, L. Xiong, Y. Xiao, J. Gao, A. R. Post, X. Jiang, and L. Ohno-Machado. "SHARE: System Design and Case Studies for Statistical Health Information Release". In: *Journal of the American Medical Informatics Association* 20.1 (2013), pp. 109–116.

[129] J. Gehrke, M. Hay, E. Lui, and R. Pass. "Crowd-Blending Privacy". In: *International Cryptology Conference (CRYPTO)*. Ed. by R. Safavi-Naini and R. Canetti. Springer, 2012.

[130] R. Hall, L. Wasserman, and A. Rinaldo. "Random Differential Privacy". arXiv preprint arXiv:1112.2680, 2013.

[131] M. Hardt and K. Talwar. "On the Geometry of Differential Privacy". In: *Symposium on Theory of Computing (STOC)*. ACM, 2010.

[132] G. Kellaris, S. Papadopoulos, and D. Papadias. "Engineering Methods for Differentially Private Histograms: Efficiency Beyond Utility". In: *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 31.2 (2019), pp. 315–328.

[133] Y. N. Kobliner, U. Stemmer, R. B. Y. Bassily, and A. G. Thakurta. "Locally Private Determination of Heavy Hitters". U.S. pat. 20180336357A1. Harvard College, University of California, Georgetown University. 2018.

[134] T. Kulkarni, G. Cormode, and D. Srivastava. "Answering Range Queries Under Local Differential Privacy". arXiv preprint arXiv:1812.10942, 2018.

[135] S. Lan, Y.-j. Wu, X.-l. Xhang, and Y. Xie. "Greedy Algorithm Based on Bucket Partitioning for Differentially Private Histogram Publication". In: *Journal of Xiamen University* (2013).

[136] J. Lei. "Differentially Private M-Estimators". In: *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2011.

[137] H. Li, Y. Dai, and X. Lin. "Efficient E-Health Data Release with Consistency Guarantee under Differential Privacy". In: *International Conference on E-Health Networking, Application & Services (HealthCom)*. IEEE, 2015.

[138] B. Li, V. Karwa, A. Slavković, and R. C. Steorts. "A Privacy Preserving Algorithm to Release Sparse High-Dimensional Histograms". In: *Journal of Privacy and Confidentiality* 8.1 (2018).

[139] X. Li, J. Yang, Z. Sun, and J. Zhang. "Differentially Private Release of the Distribution of Clustering Coefficients across Communities". In: *Security and Communication Networks* 2019 (2019), pp. 1–9.

[140] Y. Li, X. Ren, S. Yang, and X. Yang. "Impact of Prior Knowledge and Data Correlation on Privacy Leakage: A Unified Analysis". In: *IEEE Transactions on Information Forensics and Security* 14.9 (2019), pp. 2342–2357.

[141] B.-R. Lin and D. Kifer. "Information Preservation in Statistical Privacy and Bayesian Estimation of Unattributed Histograms". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2013.

[142] G. Ling, Y. Xudong, L. Zhao, M. Yong, S. Qian, and W. Fan. "Detrended Analysis Differential Privacy Protection-Based Histogram Data Release Method". Pat. CN108446568A (China). 2018.

[143] C. Luo, X. Liu, W. Xue, Y. Shen, J. Li, W. Hu, and A. X. Liu. "Predictable Privacy-Preserving Mobile Crowd Sensing: A Tale of Two Roles". In: *IEEE/ACM Transactions on Networking* 27.1 (2019), pp. 361–374.

[144] E. Naghizade, J. Bailey, L. Kulik, and E. Tanin. "Challenges of Differentially Private Release of Data Under an Open-World Assumption". In: *International Conference on Scientific and Statistical Database Management (SSDBM)*. ACM, 2017.

[145] A. Nikolov, K. Talwar, and L. Zhang. "The Geometry of Differential Privacy: The Small Database and Approximate Cases". In: *SIAM Journal on Computing* 45.2 (2016), pp. 575–616.

[146] J. Raigoza. "Differential Private-Hilbert: Data Publication Using Hilbert Curve Spatial Mapping". In: *International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2017.

[147] A. Roth. "Differential Privacy and the Fat-Shattering Dimension of Linear Queries". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. Springer, 2010, pp. 683–695.

[148] S. Shang, T. Wang, P. W. Cuff, and S. R. Kulkarni. "The Application of Differential Privacy for Rank Aggregation: Privacy and Accuracy". In: *International Conference on Information Fusion (FUSION)*. IEEE, 2014.

[149] D. B. Smith, K. Thilakarathna, and M. A. Kaafar. "More Flexible Differential Privacy: The Application of Piecewise Mixture Distributions in Query Release". arXiv preprint arXiv:1707.01189, 2017.

[150] D. Su, J. Cao, N. Li, and M. Lyu. "PrivPfC: Differentially Private Data Publication for Classification". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2018.

[151] X. Xiao, G. Bender, M. Hay, and J. Gehrke. "iReduct: Differential Privacy with Reduced Relative Errors". In: *International Conference on Management of Data (SIGMOD)*. ACM, 2011.

[152] X. Xiaoling, L. Huiyi, S. Xiujin, W. Shaoyu, and Y. Shoujian. "Histogram-Based Data Flow-Oriented Differential Privacy Publishing Method". Pat. CN105046160A (China). 2015.

[153] X. Ying, X. Wu, and Y. Wang. "On Linear Refinement of Differential Privacy-Preserving Query Answering". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Springer, 2013.

[154] L. Zhang, K. Talwar, and A. Nikolov. "Differentially Private Linear Queries on Histograms". U.S. pat. 9672364B2. Microsoft Technology Licensing LLC. 2017.

[155] T. Zhu, P. Xiong, G. Li, and W. Zhou. "Correlated Differential Privacy: Hiding Information in Non-IID Data Set". In: *IEEE Transactions on Information Forensics and Security* 10.2 (2015), pp. 229–242.

[156] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. M. Thuraisingham, and L. Sweeney. "Privacy Preserving Synthetic Data Release Using Deep Learning". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. Springer, 2018.

[157] J. M. Abowd and L. Vilhuber. "How Protective Are Synthetic Data". In: *International Conference on Privacy in Statistical Databases (PSD)*. Springer, 2008.

[158] M. Aliakbarpour, I. Diakonikolas, and R. Rubinfeld. "Differentially Private Identity and Closeness Testing of Discrete Distributions." arXiv preprint arXiv:1707.05497, 2017.

[159] M. Balog, I. Tosltikhin, and B. Schölkopf. "Differentially Private Database Release via Kernel Mean Embeddings". In: *International Conference on Machine Learning (ICML)*. ACM, 2018.

[160] A. F. Barrientos, A. Bolton, T. Balmat, J. P. Reiter, J. M. de Figueiredo, A. Machanavajjhala, Y. Chen, C. Kneifel, and M. DeLong. "Providing Access to Confidential Research Data through Synthesis and Verification: An Application to Data on Employees of the U.S. Federal Government". In: *The Annals of Applied Statistics (AOAS)* 12.2 (2018), pp. 1124–1156.

[161] A. F. Barrientos, J. P. Reiter, A. Machanavajjhala, and Y. Chen. "Differentially Private Significance Tests for Regression Coefficients". In: *Journal of Computational and Graphical Statistics* 28.2 (2018), pp. 1–24.

[162] V. Bindschaedler, R. Shokri, and C. A. Gunter. "Plausible Deniability for Privacy-Preserving Data Synthesis". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2017.

[163] A. Blum, K. Ligett, and A. Roth. "A Learning Theory Approach to Noninteractive Database Privacy". In: *Journal of the ACM* 60.2 (2013), p. 12.

[164] J. Böhler, D. Bernau, and F. Kerschbaum. "Privacy-Preserving Outlier Detection for Data Streams". In: *Conference on Data and Applications Security and Privacy (DBSec)*. IFIP WG 11.3, 2017.

[165] O. Bousquet, R. Livni, and S. Moran. "Passing Tests without Memorizing: Two Models for Fooling Discriminators". arXiv preprint arXiv:1902.03468, 2019.

[166] C. M. Bowen and F. Liu. "Differentially Private Data Synthesis Methods". arXiv preprint arXiv:1602.01063, 2016.

[167] Y. Cao, M. Yoshikawa, Y. Xiao, and L. Xiong. "Quantifying Differential Privacy under Temporal Correlations". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2017.

[168] Y. Cao, Y. Xiao, L. Xiong, and L. Bai. "PriSTE: From Location Privacy to Spatiotemporal Event Privacy." arXiv preprint arXiv:1810.09152, 2018.

[169] A.-S. Charest. "How Can We Analyze Differentially-Private Synthetic Datasets?" In: *Journal of Privacy and Confidentiality* 2.2 (2011), p. 3.

[170] L. Chen, T. Yu, and R. Chirkova. "WaveCluster with Differential Privacy". In: *International Conference on Information and Knowledge Management (CIKM)*. ACM, 2015.

[171] G. Cormode, T. Kulkarni, and D. Srivastava. "Constrained Private Mechanisms for Count Data". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2018.

[172] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. P. Vadhan. "On the Complexity of Differentially Private Data Release: Efficient Algorithms and Hardness Results". In: *Symposium on Theory of Computing (STOC)*. ACM, 2009.

[173] M. J. Elliot. "Empirical Differential Privacy: An New Method for Measuring Residual Dis-Cloure Risk in Synthetic Data". 2014.

[174] L. Fan and L. Xiong. "Adaptively Sharing Time-Series with Differential Privacy". arXiv preprint arXiv:1202.3461, 2012.

[175] S. Garfinkel. *De-Identifying Government Datasets*. Technical Report. National Institute of Standards and Technology, 2016.

[176] S. Garfinkel. *De-Identifying Government Datasets (2nd Draft)*. Technical Report. National Institute of Standards and Technology, 2016.

[177] A. Gupta, A. Roth, and J. Ullman. "Iterative Constructions and Private Data Release". In: *Theory of Cryptography Conference (TCC)*. Springer, 2012.

[178] M. A. W. Hardt. "A Study of Privacy and Fairness in Sensitive Data Analysis". PhD thesis. Princeton University, 2011.

[179] Y. Hu, H. Shen, G. Bai, and T. Wang. "Privacy-Preserving Task Allocation for Edge Computing Enhanced Mobile Crowdsensing". In: *International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP)*. Springer, 2018.

[180] B. Hu, B. Zhou, Q. Yan, A. Alim, F. Chen, and H. Zeng. "PSCluster: Differentially Private Spatial Cluster Detection for Mobile Crowdsourcing Applications". In: *Conference on Communications and Network Security (CNS)*. IEEE, 2018.

[181] J. Jordon, J. Yoon, and M. van der Schaar. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees". In: *International Conference on Learning Representations (ICLR)*. 2019.

[182] Z. Jorgensen, T. Yu, and G. Cormode. "Conservative or Liberal? Personalized Differential Privacy". In: *International Conference on Data Engineering (ICDE)*. IEEE, 2015.

[183] D. Kifer and B. R. Lin. "Towards an Axiomatization of Statistical Privacy and Utility". In: *Symposium on Principles of Database Systems (PODS)*. ACM, 2010.

[184] T. Kulkarni, G. Cormode, and D. Srivastava. "Constrained Differential Privacy for Count Data". arXiv preprint arXiv:1710.00608, 2017.

[185] J. Lee. "On Sketch Based Anonymization That Satisfies Differential Privacy Model". In: *Canadian Conference on Advances in Artificial Intelligence*. Springer, 2010.

[186] C. Li and G. Miklau. "Optimal Error of Query Sets under the Differentially-Private Matrix Mechanism". In: *International Conference on Database Theory (ICDT)*. ACM, 2013.

[187] H. Li, L. Xiong, L. Zhang, and X. Jiang. "DPSynthesizer: Differentially Private Data Synthesizer for Privacy Preserving Data Sharing". In: *International Conference on Very Large Data Bases (VLDB)*. ACM, 2014.

[188] C. Li and G. Miklau. "Lower Bounds on the Error of Query Sets Under the Differentially-Private Matrix Mechanism". In: *Theory of Computing Systems* 57.4 (2015), pp. 1159–1201.

[189] M. Li and X. Ma. "Bayesian Networks-Based Data Publishing Method Using Smooth Sensitivity". In: *International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*. IEEE, 2018.

[190] Y. Li, H. Xiao, Z. Qin, C. Miao, L. Su, J. Gao, K. Ren, and B. Ding. "Towards Differentially Private Truth Discovery for Crowd Sensing Systems". arXiv preprint arXiv:1810.04760, 2018.

[191] F. Liu. "Model-Based Differentially Private Data Synthesis". arXiv preprint arXiv:1606.08052, 2016.

[192] K.-C. Liu, C.-W. Kuo, W.-C. Liao, and P.-C. Wang. "Optimized Data De-Identification Using Multidimensional k-Anonymity". In: *International Conference On Trust, Security And Privacy In Computing and Communications/International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 2018.

[193] P.-H. Lu and C.-M. Yu. "POSTER: A Unified Framework of Differentially Private Synthetic Data Release with Generative Adversarial Network". In: *Conference on Computer and Communications Security (CCS)*. ACM, 2017.

[194] G. J. Matthews and O. Harel. "Data Confidentiality: A Review of Methods for Statistical Disclosure Limitation and Methods for Assessing Privacy". In: *Statistics Surveys* 5 (2011), pp. 1–29.

[195] D. McClure and J. P. Reiter. "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data". In: *Transactions on Data Privacy (TDP)* 5.3 (2012), pp. 535–552.

[196] D. R. McClure. "Relaxations of Differential Privacy and Risk/Utility Evaluations of Synthetic Data and Fidelity Measures". PhD thesis. Duke University, 2015.

[197] Y. Mülle, C. Clifton, and K. Böhm. "Privacy-Integrated Graph Clustering Through Differential Privacy." In: *EDBT/ICDT Workshops*. 2015.

[198] S. Neel, A. Roth, and Z. S. Wu. "How to Use Heuristics for Differential Privacy". arXiv preprint arXiv:1811.07765, 2018.

[199] M.-J. Park and H. J. Kim. "Statistical Disclosure Control for Public Microdata: Present and Future". In: *Korean Journal of Applied Statistics* 29.6 (2016), pp. 1041–1059.

[200] H. Ping, J. Stoyanovich, and B. Howe. "DataSynthesizer: Privacy-Preserving Synthetic Datasets". In: *International Conference on Scientific and Statistical Database Management (SSDBM)*. ACM, 2017.

[201] L. Rodriguez and B. Howe. "Privacy-Preserving Synthetic Datasets Over Weakly Constrained Domains". arXiv preprint arXiv:1808.07603, 2018.

[202] N. Shlomo. "Statistical Disclosure Limitation: New Directions and Challenges". In: *Journal of Privacy and Confidentiality* 8.1 (2018).

[203] J. Snoke and A. B. Slavkovic. "pMSE Mechanism: Differentially Private Synthetic Data with Maximal Distributional Similarity". In: *International Conference on Privacy in Statistical Databases (PSD)*. Springer, 2018.

[204] J. V. Snoke. "Statistical Data Privacy Methods for Increasing Research Opportunities". PhD thesis. Pennsylvania State University, 2018.

[205] A. Triastcyn and B. Faltings. "Generating Differentially Private Datasets Using GANs". arXiv preprint arXiv:1803.03148, 2018.

[206] J. R. Ullman. "Privacy and the Complexity of Simple Queries". PhD thesis. Harvard, 2013.

[207] L. Vilhuber, J. M. Abowd, and J. P. Reiter. "Synthetic Establishment Microdata around the World". In: *Statistical Journal of the IAOS* 32.1 (2016), pp. 65–68.

[208] J. Wang, Z. Cai, Y. Li, D. Yang, J. Li, and H. Gao. "Protecting Query Privacy with Differentially Private K-Anonymity in Location-Based Services". In: *Personal and Ubiquitous Computing* 22.3 (2018), pp. 453–469.

[209] Z. Wang, Y. Zhu, and X. Zhou. "A Data Publishing System Based on Privacy Preservation". In: *International Conference on Database Systems for Advanced Applications (DASFAA)*. Springer, 2019.

[210] B. Weggenmann and F. Kerschbaum. "SynTF: Synthetic and Differentially Private Term Frequency Vectors for Privacy-Preserving Text Mining". In: *International Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 2018.

[211] C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, and K. Ren. "GANobfuscator: Mitigating Information Leakage under GAN via Differential Privacy". In: *IEEE Transactions on Information Forensics and Security* 14.9 (2019), pp. 2358–2371.

[212] H. Yu. "Differentially Private Verification ofPredictions from Synthetic Data". Master thesis. Duke University, 2017.

[213] J. Zhang. "Algorithms for Synthetic Data Release under Differential Privacy". PhD thesis. Nanyang Technological University, 2016.

[214] X. Zhang, S. Ji, and T. Wang. "Differentially Private Releasing via Deep Generative Model". arXiv preprint arXiv:1801.01594, 2018.

[215] S. Zhou, K. Ligett, and L. A. Wasserman. "Differential Privacy with Compression". In: *International Symposium on Information Theory (ISIT)*. IEEE, 2009.

A Excluded Papers

Query 1

Table 12: Excluded papers from query 1 (focusing on histograms), and the corresponding exclusion criteria.

Citation	Exclusion Criteria
Balcer and Vadhan [115]	2
Bassily and Smith [102]	6
Benkhelif [116]	9, 10
Bhowmick et al. [117]	7
Bowen and Liu [118]	8
Bowen and Liu [119]	8
Chaudhuri and Vinterbo [120]	5
Cyphers and Veeramachaneni [121]	5
Eugenio and Liu [122]	5
Fanaeepour et al. [123]	1
Fanaeepour and Rubinstein [124]	4
Fanaeepour and Rubinstein [125]	2
Fei et al. [126]	7
Foote et al. [127]	5
Gardner et al. [128]	2
Gehrke et al. [129]	3
Hall et al. [130]	3
Hardt and Rothblum [81]	1, 2, 6
Hardt and Talwar [131]	6
Kellaris et al. [132]	2
Kobliner et al. [133]	7
Kulkarni et al. [134]	9
Lan et al. [135]	10
Lei [136]	5
Li et al. [78]	6
Li et al. [8]	2
Li et al. [137]	2
Li et al. [138]	2
Li et al. [139]	5
Li et al. [140]	1, 2, 5
Lin and Kifer [141]	1, 2, 5
Ling et al. [142]	7
Luo et al. [143]	1, 2, 3
Meng et al. [9]	2
Naghizade et al. [144]	1
Nikolov et al. [145]	6
Raigoza [146]	2, 6, 9
Roth [147]	6

Continued on next page

Table 12 – *Continued from previous page*

Citation	**Exclusion Criteria**
Shang et al. [148]	2
Smith et al. [149]	1
Su et al. [150]	5
Xiao et al. [151]	3
Xiaoling et al. [152]	7
Ying et al. [153]	2, 6
Zhang et al. [154]	7
Zhu et al. [155]	2

Query 2

Table 13: Excluded papers from query 2 (focusing on synthetic data), and the corresponding exclusion criteria.

Citation	Exclusion Criteria
Abay et al. [156]	2
Abowd and Vilhuber [157]	1
Aliakbarpour et al. [158]	1
Balog et al. [159]	2, 6
Barak et al. [74]	6
Barrientos et al. [160]	1
Barrientos et al. [161]	1
Bindschaedler et al. [162]	3
Blum et al. [65]	2
Blum et al. [163]	2
Böhler et al. [164]	4
Bousquet et al. [165]	1
Bowen and Liu [7]	2,6
Bowen and Liu [166]	8
Bowen and Liu [119]	8
Cao et al. [167]	1, 2
Cao et al. [168]	1
Charest [169]	1
Chen et al. [170]	1
Cormode et al. [171]	1, 2
Dwork et al. [172]	2, 6
Elliot [173]	8
Fan and Xiong [92]	1
Fan and Xiong [174]	8
Garfinkel [175]	1
Garfinkel [176]	1
Gehrke et al. [129]	2
Gupta et al. [177]	2
Hardt [178]	9
Hu et al. [179]	1
Hu et al. [180]	1
Jordon et al. [181]	5
Jorgensen et al. [182]	1, 2
Kifer and Lin [183]	1
Kulkarni et al. [184]	8
Lee [185]	2
Li and Miklau [186]	1
Li et al. [187]	2
Li and Miklau [188]	1, 2
Li et al. [8]	2,6

Continued on next page

Table 13 – *Continued from previous page*

Citation	Exclusion Criteria
Li and Ma [189]	4
Li et al. [138]	2
Li et al. [190]	4
Liu [191]	1
Liu et al. [192]	2
Lu and Yu [193]	2
Machanavajjhala et al. [68]	3, 6
Matthews and Harel [194]	2, 6
McClure and Reiter [195]	1
McClure [196]	9
Mülle et al. [197]	1
Neel et al. [198]	6
Park and Kim [199]	10
Ping et al. [200]	2
Rodriguez and Howe [201]	2
Shlomo [202]	1
Snoke and Slavkovic [203]	2
Snoke [204]	9
Triastcyn and Faltings [205]	2
Ullman [206]	9
Vilhuber et al. [207]	1
Wang et al. [208]	1, 3, 5
Wang et al. [209]	2
Weggenmann and Kerschbaum [210]	1
Xu et al. [211]	1
Yu [212]	9
Zhang [213]	9
Zhang et al. [214]	5
Zhou et al. [215]	1